New developments in deep learning for natural language processing and explanation

Andrea Simeri

Supervisor: Prof. Andrea Tagarelli

Program Coordinator: Prof. Giancarlo Fortino

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica (DIMES)
Università della Calabria

This dissertation is submitted for the degree of
Doctor of Philosophy
January 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. In particular, this dissertation is based on the following published research works:

- [97] Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code
- [47] The Italian civil code network analysis
- [96] LamBERTa: Law Article Mining Based on Bert Architecture for the Italian Civil Code
- [48] LawNet-Viz: A Web-Based System to Visually Explore Networks of Law Article References
- [90] Exploring domain and task adaptation of LamBERTa models for article retrieval on the Italian Civil Code
- [91] GDPR article retrieval based on domain-adaptive and task-adaptive legal pre-trained language models

Andrea Simeri

January 2024

Abstract

This PhD thesis introduces LamBERTa, a pioneering BERT-based language understanding framework tailored for law article retrieval as a prediction task. LamBERTa addresses the challenges posed by a multi-class classification scenario featuring hundreds or thousands of classes and minimal per-class training instances, generated in an unsupervised manner. The study explores domain-adaptation coupled with task-adaptation to train LamBERTa models for the Italian Civil Code (ICC) article prediction, incorporating a recently defined Italian legal BERT and investigating the impact of enhancing the tokenizer with terms from the target legal corpus.

The work's contributions extend to the exploration of LamBERTa models' explainability, employing model-level and instance-level explainability techniques. While initially focused on the Italian Civil Code, the LamBERTa architecture proves adaptable to other law code corpora such as the General Data Protection Regulation (GDPR), further powered by self-supervised task-adaptive pre-training stages, with or without data enrichment based on recitals, marking the first comprehensive study of domain-adaptive and task-adaptive legal pre-trained language models for GDPR article retrieval.

Furthermore, the thesis conducts a network analysis and mining study, presenting the first examination of citation networks inferred from ICC articles. Insights into structural features reveal hidden patterns, facilitating new interpretations and studies of the ICC. This methodology is adaptable to other civil law code systems organized similarly to the ICC, supporting tasks like statute law retrieval, entailment, and question answering. The study also introduces LawNet-Viz, a web-based tool for modeling, analyzing, and visualizing law reference networks, aiming to assist legal experts and citizens in legal information processing and understanding.

Finally, the thesis outlines the TrustSearch project's ambitious goal to develop an AI tool that enhances the search experience promoting critical thinking. Two stance detection approaches, Definition-based and User-driven, are presented, showcasing the effectiveness of AI in transforming the search and data discovery experience.

In particular, Chapter 1 introduces the motivations behind this research work, then it overviews recent works that address legal classification and retrieval problems based on deep

learning methods and network analysis techniques as well as entailment resolution using Transformers and LLMs. Chapter 2 provides the reader with essential information about the research topic and it serves to establish the context around The Italian Civil Code (ICC) and The General Data Protection Regulation (GDPR). Chapter 3 presents our proposed framework for the civil-law article retrieval problem named as LamBERTa with the application on the ICC corpus. Within it, we investigate the performance of the models through an extensive quantitative and qualitative set of experiments. Furthermore, we study the effects of domain adaptive pre-training strategies and task adaptive training strategies and we scrutinize the concept of Explainability leveraging on the use of different techniques. Chapter 4 presents the LamBERTa models applied on the General Data Protection Regulation (GDPR) corpus, also powering them through self-supervised task-adaptive pre-training schemes. Chapter 5 explores the Italian Civil Code (ICC) from an unprecedented perspective based on network analysis. Furthermore, we present LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. Chapter 6 is based on the research I carried out during the PhD visiting period abroad at (UPV) University Polytechnic of Valencia, Spain, where I actively participated in the TrustSearch project, a research project funded by the European Union, developing an AI tool that offers a new search experience to promote critical thinking. In this regard, we developed two different approaches, the first one which is based on the state of art LLMs prompt engineer, the second one consists is an application of an encoder-decoder Transformer-based model for the stance detection leveraging on the entailment task. Chapter 7 summarizes findings and discusses future studies which could extend the research.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Artificial Intelligence (AI) and in particular NLP is increasingly used in crtitical domains as the the legal domain, which finds main motivations in the huge amount of information produced and in the involvement of different actors, such as legal professionals, law courts, legislators, law firms, and even citizens [95]- and in the health domain for instance with application in the case of mental issues recognition [36].

The general purpose of law search is to recognize legal authorities that are relevant to a question expressing a legal matter [24]. The interpretative uncertainty in law, particularly that related to the jurisprudential type which is capable of directly affecting citizens, has prompted many to model law search as a prediction problem [24]. Ultimately, this would allow lawyers and legal practitioners to explore the possibility of predicting the outcome of a judgment (e.g., the probable sentence relating to a specific case), through the aid of computational methods, also sometimes referred to as *predictive justice* [101]. Predictive justice is currently being developed, to a prevalent extent, following a statistical-jurisprudential approach: the jurisprudential precedents are verified and future decisions are predicted on their basis. However, as stated by legal professionals [101], several reasons point out that this approach should not be preferred, because of its limited scope only to cases in which there are numerous precedents, so as to exclude unprecedented cases relating to new regulations, not yet subject to stratified jurisprudential guidelines. In fact, the jurisprudential approach is not in line with any *civil law* system — which is adopted in most European Union states and non-Anglophone countries in the world — with the consequence of a high risk of fallacy (i.e., repetition of errors based on precedents) and risk of standardization (i.e., if a lawsuit is contrary to many precedents, then no one will propose such a lawsuit).

Clearly, from a major perspective as a *data-driven artificial-intelligence* task, predicting judicial decisions is carried out exclusively based on the legal corpora available and the selected algorithms to use, as remarked by Medvedeva et al. [65]. Moreover, as also witnessed by an increased interest from the artificial intelligence and law research community, a key perspective in legal analysis and problem solving lays on the opportunities given by advanced, data-driven computational approaches based on natural language processing (NLP), data mining, and machine learning disciplines [22].

Predictive tasks in legal information systems have often been addressed as text classification problems, ranging from case classification and legal judgment prediction [69, 58, 57, 4, 93, 102, 65], to legislation norm classification [12], and statute prediction [60]. Early studies have focused on statistical textual features and machine learning methods, then the progress of *deep learning* methods for text classification [35, 34] has prompted the development of deep neural network frameworks, such as recurrent neural networks, for single-task learning (e.g., charge prediction [63, 109], sentence modality classification [76, 18], legal question answering [27]) or even multi-task learning (e.g., [107, 117]).

More recently, *deep pre-trained language models*, particularly the Bidirectional Encoder Representations from Transformers (BERT) [25], have emerged showing outstanding effectiveness in several NLP tasks. Thanks to their ability to learn a contextual language understanding model, these models overcome the need for feature engineering (upon which classic, sparse vectorial representation models rely). Nonetheless, since they are originally trained from generic domain corpora, they should not be directly applied to a specific domain corpus, as the distributional representation (embeddings) of their lexical units may significantly shift from the nuances and peculiarities expressed in domain-specific texts — and this certainly holds for the legal domain as well, where interpreting and relating documents is particularly challenging.

Developing BERT models for legal texts has very recently attracted increased attention, mostly concerning classification problems (e.g., [81, 17, 88, 89, 20, 111, 73]). Our research falls into this context, as we propose a BERT-based framework for law article retrieval based on civil-law-based corpora. More specifically, as we wanted to benefit from the essential consultation provided by law professionals in our country, our proposed framework is completely specified using the Italian Civil Code (ICC) as the target legal corpus. Notably, only few works have been developed for Italian BERT-based models, such as a retrained BERT for various NLP tasks on Italian tweets [79], and a BERT-based masked-language model for spell correction [80]; however, *no study leveraging BERT for the Italian civil-law corpus has been proposed so far*.

Extending the focus from national level to the international level, The General Data Protection Regulation (GDPR) stands as one of the most significant legal frameworks for data protection and privacy in recent years. Enforced by the European Union (EU) since May 2018, the GDPR has garnered global attention due to its wide-reaching impact on businesses, organizations, and individuals, transcending geographical boundaries. Automating the search for GDPR information is demanding to address the need for efficient access to the GDPR contents. By employing advanced NLP technologies, the process of accessing and understanding GDPR provisions can be greatly facilitated. In this regard, *pre-trained language models* (PLMs) such as BERT and GPT like models are the most helpful and attractive tools, given their widely known remarkable capabilities in various NLP tasks, including text classification, question answering, and document retrieval. However, despite the potential benefits of leveraging PLMs for GDPR article retrieval, we are not aware of PLMs that have been specifically trained on GDPR texts to date. For these reasons, we decided to apply our BERT-based framework for law article retrieval even on the GDPR corpus.

Research in artificial intelligence and law has traditional focused on a number of problems that are typically addressed by natural language processing and machine learning models and methods. Despite a relative gap with respect to the network science discipline, a few works have been proposed to investigate the complexity in a legal corpus domain by leveraging network analysis methods. Therefore, we decided to explore the Italian Civil Code (ICC) from an unprecedented perspective based on network analysis. To the best of our knowledge, the Italian Civil Code has not been studied through its article-reference-based structure so far. In this thesis. we perform a study of the networks that can be inferred from the ICC corpus based on article references. This allows us to extend our scope beyond the canonical exploration boundaries of the legal domain, by providing new insights on the interpretation and evaluation of the Italian civil law.

Despite such important contributions in visual legal analytics and knowledge graph research, to the best of our knowledge there is still a lack of systems integrating visual and network analysis with recently developed, advanced methods in natural language processing, particularly the deep contextualized language models such as BERT [25]. The latter is instead a key point in the development of our LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. LawNet-Viz is designed to support legal research tasks and help legal professionals as well as laymen visually exploring the article connections built upon the explicit law references detected in the article contents.

Finally, we move our attention in the field of NLI Natural Language Inference, in particular on the entailment task, and how to use it to develop users' critical thinking. Indeed, in the last decade we appreciated an increasing mistrust in all information sources. Following the Edelman Trust Barometer 2021[1], we see that trust in information is at record lows. Also, confidence in search engines (-9%) and traditional media (-12%) has decreased in the last two years. According to this, the Next generation of the Internet should comprise a change in the way we experience search and discover data and resources on the internet and web. A new generation of search engines that promotes critical thinking must be developed. In particular, stance detection is an emerging opinion mining paradigm for various social and political applications. Recently there has been a growing research interest in detecting automatically the stance from opinionated texts, especially if about controversial topics (e.g. vaccines, climate change, immigration, etc.). The recent survey in [2] presents an exhaustive review of stance detection techniques, the different types of targets, the features set used, and the best performing machine learning / deep learning approaches.

1.2 Contributions

In this section we present our main contributions in the research. For clarity, they are divided by chapters. Our main contributions in Chapter 3 are summarized as follows:

- We push forward research on law document analysis for civil law systems, focusing on the modeling, learning and understanding of logically coherent corpora of law articles, using the Italian Civil Code as case in point.
- We study the law article retrieval task as a prediction problem based on the deep machine learning paradigm. More specifically, following the lastest advances in research on deep neural network models for text data, we propose a deep pre-trained contextualized language model framework, named LamBERTa (Law article mining based on BERT architecture). LamBERTa is designed to fine-tune an Italian pre-trained BERT on the ICC corpora for law article retrieval as prediction, i.e., given a natural language query, predict the most relevant ICC article(s).
- Notably, we deal with a very challenging prediction task, which is characterized not only by a high number (i.e., hundreds) of classes — as many as the number of articles — but also by the issues that arise from the need for building suitable training sets given the lack of test query benchmarks for Italian legal article retrieval/prediction

[1] https://www.edelman.com/trust/archive

tasks. This also leads to coping with few-shot learning issues (i.e., learning models to predict the correct class of instances when a small amount of examples are available in the training dataset), which has been recognized as one of the so-called *extreme classification* scenarios [9, 19]. We design our LamBERTa framework to solve such issues based on different schemes of *unsupervised training-instance labeling* that we originally define for the ICC corpus, although they can easily be generalized to other law code systems.

- We address one crucial aspect that typically arises in deep/machine learning models, namely *explainability*, which is clearly of interest also in artificial intelligence and law (e.g., [14, 38]). In this regard, we investigate explainability of our LamBERTa models through two different approaches. The first one focusing on the understanding of how they form complex relationships between the textual tokens. The second one using LIME - Local Interpretable Model-Agnostic Explanations [86] o explain the behavior of the underlying classifier "around" a query instance. We further provide insights into the patterns generated by LamBERTa models through a visual exploratory analysis of the learned representation embeddings.

- We present an extensive, quantitative experimental analysis of LamBERTa models by considering:

  – six different types of test queries, which vary by originating source, length and lexical characteristics, and include *comments* about the ICC articles as well as *case law decisions* from the civil section of the Italian Court of Cassation that contain significant jurisprudential sentences associated with the ICC articles;

  – *single-label* evaluation as well as *multi-label* evaluation tasks;

  – different sets of assessment criteria.

- We investigate the behavior of LamBERTa models when fine-tuning a legal Italian BERT rather than a general-domain Italian BERT;

- We present results about the impact of injecting out-of-vocabulary legal terms into LamBERTa models during the fine-tuning stage using different strategies for tokens' initialization;

- We investigate the effect of domain-adaptation and task-adaptation of a pre-trained Italian BERT model on the performance.

The obtained results have shown the effectiveness of LamBERTa, and its superiority against (i) widely used deep-learning text classifiers that have been tested on our different query sets for the article prediction tasks, and against (ii) a few-shot learner conceived for an attribute-aware prediction task that we have newly designed based on the availability of ICC metadata.

Our main contributions in Chapter 4 are summarized as follows:

- We train our LamBERTa models for the task of GDPR article retrieval. More specifically, we train and fine-tune a pool of BERT models for a sequence classification task on the GDPR articles, with or without data enrichment based on the GDPR recitals.
- We conduct extensive experiments and evaluations including not only the general domain (i.e., base) BERT but also the *legal BERT* models in [20], particularly the from-scratch pre-trained and further pre-trained versions.
- We propose two self-supervised task-adaptive pre-training strategies, namely *Related Sentence Prediction* and *Multiple Choice Answering*, which show key advantages on different query sets.

Our main contributions in Chapter 5 are summarized as follows:

- The ICC articles are not commonly available as hypertexts. Therefore, we develop a text processing method to identify and extract article references that are valid w.r.t. the ICC, hence discarding references to legal sources that are external to the ICC.
- Based on the article-reference relation, we define two network models by differentiating at level of single book or entire ICC corpus.
- We perform an analysis of the book-induced and corpus networks in order to unveil main structural features of such networks and interesting patterns underlying the article references within and across the books of the ICC. Our analysis of the networks is twofold, as we consider macro-scale and meso-scale properties of the networks.
- Upon the outcomes of our performed task of community detection, we also investigate on relations between the formation of communities and the topic coherence within and across books of the ICC. Our findings reveal useful indicators that may help legal experts and practitioners integrate their knowledge from a novel perspective that lays on a unifying, mesoscopic organization of the network of article relations spanning over the whole ICC.

- We contribute with a first study of the networks that can be inferred from the ICC corpus based on article references;
- We present LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus.
- LawNet-Viz allows the user to refine a network according to the semantic similarity of linked articles, which is captured based on sparse vectorial representations, topic models, word embeddings or deep contextualized embeddings coming from LamBERTa or other models.

Our main contributions in Chapter 6 are summarized as follows:

- We propose a stance detection approach in order to develop users' critical thinking in a new generation of search engines.
- We realized a working demo with two possible stance detection approaches (i) Definition-based approach which leverages on the knowledge expressed by LLMs (ii) User-driven approach which leverages on state of the art Transformer model pre-trained on the entailment task.

1.3 Related works

Artificial Intelligence & Law. Our work belongs to the corpus of studies that reflect the recent revival of interest in the role that machine learning, particularly deep neural network models, can take on artificial intelligence applications for text data in a variety of domains, including the legal one. In this regard, here we overview recent research works that employ deep learning methods for addressing computational problems in the legal domain, with a focus on classification and retrieval tasks. Note that the latter are major categories for the data-driven legal analysis literature review, along with entailment and information extraction based on NLP approaches (e.g., named entity recognition, relation extraction, tagging), as extensively studied by Chalkidis and Kampas in [21], to which we refer the interested reader for a broader overview.

Most existing works on deep-learning-based law analysis exploit recurrent neural network models (RNNs) and convolutional neural networks (CNNs), along with the classic multi-layer perceptron (MLP). For instance, O'Neill et al. [76] utilize all the above methods for classifying deontic modalities in regulatory texts, demonstrating superiority of neural network models over competitive classifiers including ensemble-based decision tree and

largest margin classifiers. Focusing on obligation and prohibition extraction as a particular case of deontic sentence classification, Chalkidis et al. [18] show the benefits of employing a hierarchical attention-based bidirectional LSTM model that considers both the sequence of words in each sentence and the sequence of sentences. Branting et al. [15] consider administrative adjudication prediction in motion-rulings, Board of Veterans Appeals issue decisions, and World Intellectual Property Organization domain name dispute decisions. In this regard, three approaches for prediction are evaluated: maximum entropy over token n-grams, SVM over token n-grams, and a hierarchical attention network [108] applied to the full text. While no absolute winner was observed, the study highlights the benefit of using feature weights or network attention weights from these predictive models to identify salient phrases in motions or contentions and case facts. Nguyen et al. [74, 75] propose several approaches to train long short term memory (LSTMs) models and conditional random field (CRF) models for the problem of identifying two key portions of legal documents, i.e., requisite and effectuation segments, with evaluation on Japanese civil code and Japanese National Pension Law dataset. In [21] by Chalkidis and Kampas, a major contribution is the development of word2vec skip-gram embeddings trained on large legal corpora (mostly from European, UK, and US legislations).

Note that our work is clearly different from the aforementioned ones, since they not only focus on legal corpora other than Italian civil law articles but also they consider machine learning and neural network models that do not exploit the same ability as pre-trained deep language models.

To address the problem of predicting the final charges according to the fact descriptions in criminal cases, Hu et al. [40] propose to exploit a set of categorical attributes to discriminate among charges (e.g., violence, death, profit purpose, buying and selling). By leveraging these annotations of charges based on representative attributes, the proposed learning framework aims to predict attributes and charges of a case simultaneously. An attribute attention mechanism is first applied to select factual information from facts that are relevant to each particular attribute, so to generate attribute-aware fact representations that can be used to predict the label of an attribute, under a binary classification task. Then, for the task of charge prediction, the attribute-aware fact representations aggregated by average pooling are also concatenated with the attribute-free fact representations produced by a conventional LSTM neural network. The training objective is twofold, as it minimizes the cross-entropy loss of charge prediction and the cross-entropy loss of attribute prediction.

It should be noticed that the above study was especially designed to deal with the typical imbalance of the case numbers of various charges as well as to distinguish related or "confusing" charges. In particular, the first aspect corresponds to a challenge of insufficient

training data for some charges, as there are indeed charges with limited cases. This appears to be in analogy with the few-shot learning scenario in law article prediction; therefore, in Section 3.3.3, we shall present a comparative evaluation stage with the method in [40] adapted for the ICC law article prediction task.

Also in the context of legal judgment prediction, Li et al. [55] propose a multichannel attention-based neural network model, dubbed MANN, that exploits not only the case facts but also information on the defendant persona, such as traits that determine the criminal liability (e.g., age, health condition, mental status) and criminal records. A two-tier structure is used to empower attention-based sequence encoders to hierarchically model the semantic interactions from different parts of case description. Results on datasets of criminal cases in mainland China have shown improvements over other neural network models for judgment prediction, althugh MANN may suffer from the imbalanced classes of prison terms and cannot deal with criminal cases with multiple defendants. More recently, Gan et al. [33] have proposed to inject legal knowledge into a neural network model to improve performance and interpretability of legal judgment prediction. The key idea is to model declarative legal knowledge as a set of first-order logic rules and integrate these logic rules into a co-attention network based model (i.e., bidirectional information flows between facts and claims) in an end-to-end way. The method has been evaluated on a collection of private loan law cases, where each instance in the dataset consists of a fact description and the plaintiff's multiple claims, demonstrating some advantage over AutoJudge [61], which models the interactions between claims and fact descriptions via pair-wise attention in a judgment prediction task.

The above two works are distant from ours, not only in terms of the target corpora and addressed problems but also since we do not use any type of information other than the text of the articles, nor any injected knowledge base.

In the last few years, the Competition on Legal Information Extraction/ Entailment (COLIEE) has been an important venue for displaying studies focused on case/statute law retrieval, entailment, and question answering. In most works appeared in the most recent COLIEE editions, the observed trend is to tackle the retrieval task by using CNNs and RNNs in the entailment phase, possibly in combination of additional features produced by applying classic term relevance weighting methods (e.g., TF-IDF, BM25) or statistical topic models (e.g., Latent Dirichlet Allocation). For instance, Kim et al. [44] propose a binary CNN-based classifier model for answering to the legal queries in the entailment phase. The entailment model introduced by Morimoto et al. [67] is instead based on MLP incorporating the attention mechanism. Nanda et al. [70] adopt a combination of partial string matching and topic clustering for the retrieval task, while they combine LSTM and CNN models for

the entailment phase. Do et al. [27] propose a CNN binary model with additional TF-IDF and statistical latent semantic features.

The aforementioned studies differ from ours as they mostly focus on CNN and RNN based neural network models, which are indeed used as competing methods against our proposed deep-pretrained language model framework (cf. Section 3.3.3).

Exploiting BERT for law classification tasks has recently attracted much attention. Besides a study on Japanese legal term correction proposed by Yamakoshi et al. [106], a few very recent works address sentence-pair classification problems in legal information retrieval and entailment scenarios. Rabelo et al. [81] propose to combine similarity based features and BERT fine-tuned to the task of case law entailment on the data provided by the Competition on Legal Information Extraction/Entailment (COLIEE), where the input is an entailed fragment from a case coupled with a candidate entailing paragraph from a noticed case. Sanchez et al. [88] employ BERT in its regression form to learn complex relevance criteria to support legal search over news articles. Specifically, the input consists of a query-document pair, and the output is a predicted relevance score. Results have shown that BERT trained either on a combined title and summary field of documents or on the documents' contents outperform a learning-to-rank approach based on LambdaMART equipped with features engineered upon three groups of relevance criteria, namely topical relevance, factual information, and language quality. However, in legal case retrieval, the query case is typically much longer and more complex than common keyword queries, and the definition of relevance between a query case and a supporting case could be beyond general topical relevance, which makes it difficult to build a large-scale case retrieval dataset. To address this challenge, Shao et al. [89] propose a BERT framework to model semantic relationships to infer the relevance between two cases by aggregating paragraph-level dependencies. To this purpose, the BERT model is fine-tuned with a relatively small-scale case law entailment dataset to adapt it to the legal scenario. Experiments conducted on the benchmark of the relevant case retrieval task in COLIEE 2019 have shown effectiveness of the proposed BERT model.

We notice that the above two works require to classify query-document pairs (i.e., pairs of query and news article, in [88], or pairs of case law documents, in [89]), whereas our models are trained by using articles only. At the time of submission of this article, we also become aware of a small bunch of works presented at COLIEE-2020 that compete for the statute law retrieval and question answering (statute entailment) tasks[2] using BERT [82]. In particular, for the statute law retrieval task, the goal is to read a legal bar exam question and retrieve a subset of Japanese civil code articles to judge whether the question is entailed or not. The BERT-based approach has shown to improve overall retrieval performance, although there

[2]https://sites.ualberta.ca/ rabelo/COLIEE2020/.

are still numbers of questions that are difficult to retrieve by BERT too. At COLIEE-2021, which was held in June 2021,[3] there has been an increased attention and development of BERT-based methods to address the statute and case law processing tasks [111, 73].

Again, it should be noted that the above works at COLIEE Competitions assume that the training data are questions and relevant article pairs, whereas our training data instances are derived from articles only. Nonetheless, we recognize that some of the techniques introduced at COLIEE-2021, such as deploying weighted aggregation on models' predictions and the iterated self-labeled and fine-tuning process, are worthy of investigation and we shall delve into them in our future research.

In addition to the COLIEE Competitions, it is worth mentioning the study by Chalkidis et al. [17], which provides a threefold contribution. They release a new dataset of cases from the European Court of Human Rights (ECHR) for legal judgment prediction, which is larger (about 11.5k cases) than earlier datasets, and such that each case along with its list of facts is mapped to articles violated (if any) and is assigned an ECHR importance score. The dataset is used to evaluate a selection of neural network models, for different tasks, namely binary classification (i.e., whether a case violates a human rights article or not), multi-label classification (i.e., which types of violation, if any), and case importance detection. Results have shown that the neural network models outperform an SVM model with bag-of-words features, which was previously used in related work such as [4]. Moreover, a hierarchical version of BERT, dubbed HIER-BERT, is proposed to overcome the BERT's maximum length limitation, by first generating fact embeddings and then using them through a self-attention mechanism to produce case embeddings, similarly to a hierarchical attention network model [108].

The latter aspect on the use of a hierarchical attention mechanism, especially when integrated into BERT, is very interesting and useful on long legal documents, such as case law documents, to improve the performance of pre-trained language models like BERT that are designed with constraints on the tokenized text length. Nonetheless, as we shall discuss later in Section 3.1.4, this contingency does not represent an issue in the setting of our proposed framework, due not only to the characteristic length of ICC articles but also to our designed schemes of unsupervised training-instance labeling.

Much more recently, a new contribution to the Italian legal domain has been offered by the release of the first Italian BERT pre-trained on legal corpora, named ITALIAN-LEGAL-BERT [56].

[3] https://sites.ualberta.ca/ rabelo/COLIEE2021/.

Deep Learning & GDPR. Extending the focus from national level to the international level, in the European Union one of the most important law corpus is the GDPR. Most existing approaches have focused on checking the GDPR *compliance* of privacy policies.

[98] proposes a conceptual model for characterizing the content of privacy policies in terms of information elements that one can expect to find in them (e.g., controller's identity and contact). Based on named entity recognition and Glove word embeddings, such information elements are extracted and used to train a SVM model for a task of multi-label classification. The approach in [39] distinguishes between coarse-grained and fine-grained practices based on the OPP-15 taxonomy, and model them as a directed acyclic graph with a three level structure (i.e., categories, attributes and values) so that the extraction of data practices is treated as a hierarchical multi-label classification task converted into two text-to-text tasks, one for each level of the label hierarchy, based on a T5 model. The extracted information are fed into a rule-based system that encodes the GDPR articles 13 and 14 under the supervision of legal experts in accord with the OPP-115 taxonomy.

[29] identifies four types of regulatory entities within policies of web services, which are used to define 16 classes. A BiLSTM is trained for a multi-class classification task, and a BERT summarizer is applied prior to the evaluation of context similarity for adhering vs. non-adhering policies. [3] exploits a supervised variant of Latent Dirichlet Allocation (i.e., Labeled LDA) to model topics associated with various types of violations of GDPR articles within real privacy incidents. The most relevant words associated with the topics induced by Labeled LDA are used to augment the training instances from the CMS.Law GDPR Enforcement Tracker database[4] to learn an LSTM classifier at GDPR article level. [85] leverages FastText word embeddings and a CNN model to predict privacy disclosure requirements according to GDPR articles 13 and 14. [5] is concerned with compliance of data processing agreements (DPAs), i.e., legally binding agreements that regulate the data processing activities according to GDPR. In relation to a predefined set of 45 requirements extracted from the GDPR provisions relevant to DPA, the proposed approach aims to assess whether a DPA is GDPR compliant, by comparing semantic-role-based representations of DPAs against predefined representations of the requirements. Also, the approach further provides recommendations about missing information in the DPA. In the context of GDPR compliance concerning the Italian Public Administration, [62] proposes a framework to detect security breaches related to unlawful disclosure of health information in public documents. Personally identifiable information as named entities are extracted and used to feed a machine learning classifier (e.g., SVM, XGBoost) for a binary classification task (i.e., compliant or non-compliant).

[4] https://www.enforcementtracker.com/

In comparison to the aforementioned studies, our work aims to perform a similarity search of a generic text (e.g., a policy, a DPA or an incident) with GDPR, through a full unsupervised article retrieval task. Although the most important requirements to satisfy are mentioned in a few of articles used by competitors (e.g., article 13 and article 14), there are other articles containing fundamental GDPR requirements, for example as reported in [39] the data processing should complies with the data protection rules (e.g, fair and transparent processing of Art. 5, and lawfulness of the processing of Art. 6). Furthermore, a privacy policy must specify the legal basis of the processing according to Article 13.1(c), but it must also demonstrate that this legal basis complies with Article 6 requirements on the lawfulness of the processing (e.g., if the legal basis is consent, the controller must ensure that consent has been given for one or more specific purposes.). In this regard, it should be noted that article retrieval task, can be useful for other downstream tasks as checking compliance and completeness. Through extensive experiments conducted on large-scale datasets, we demonstrate the effectiveness of our approach, achieving state-of-the-art performance w.r.t. standard metrics. Additionally, our method offers improved scalability, making it applicable to real-world scenarios. It for a fact, these tools should be treated as a support to authorities in order to look out anomaly and possible offenders as well as to stakeholders in order to identify areas where organizations may need to improve their compliance with GDPR.

Network Analysis & Law. Research in artificial intelligence and law has traditional focused on a number of problems that are typically addressed by natural language processing and machine learning models and methods. Despite a relative gap with respect to the network science discipline, a few works have been proposed to investigate the complexity in a legal corpus domain by leveraging network analysis methods. For instance, Fowler et al. [32] use a network-based representation to characterize the most important precedents at the U.S. Supreme Court. Moser et al. [68] study the structural properties of the citation networks inferred from Austrian Supreme Court decisions. Koniaris et al. [46] model the legislation network of the European Union law, and explore its topological structure and evolution over time, also evaluating its resilience. Mazzega et al. [64] study the network of French Legal codes. Moreover, Zhang et al. [113] propose a software tool to visually explore semantic-based citation networks to ease the analysis of the relations and the evolution of legal issues.

To the best of our knowledge, the Italian Civil Code has not been studied through its article-reference-based structure so far. Therefore, our main goal in this work is to contribute with a study of the networks that can be inferred from the ICC corpus based on article references. This allows us to extend our scope beyond the canonical exploration boundaries

of the legal domain, by providing new insights on the interpretation and evaluation of the Italian civil law.

Law information retrieval and related tasks have often benefited from the application of network science methodologies [104, 105, 50, 94, 66]. In [24], law search is modeled as a strategic-predictive process, where the search space follows a network-based structure. In [87], an information extraction framework is proposed to build a dynamic legislation network from legal documents.

A variety of legal sources and contexts have also been covered. For instance, network-based modeling has been utilized for the study of Courts' decisions, such as in [13] for the extraction of legal indicators from the French court of appeal judgments, in [32] to characterize the most legally relevant precedents at the U.S. Supreme Court, and in [68] to study the structural properties of the citation networks from the Austrian Supreme Court decisions. Further studies of network analysis have focused on the European Union law's legislation [46], the French Legal codes [64], the Italian Constitutional Court [1], and the Italian Civil Code [47].

As concerns visual legal analytics, the research tool proposed in [113] builds and visualizes citation networks to study legal issues in case law documents. EUCaseNet [51] is a toolkit that is comprised of a set of visualization modules and provides different analysis features that can be used by domain experts to explore on the fly the European case law corpus. Also, the Knowlex web application [52] is designed to explore and analyze legal documents from multiple sources (norms, case law, legal literature), by means of two main visual functionalities, namely interactive node graph and zoomable treemap showing the topics, the evolution, and the dimension of the legal literature settled over the years around the norm of interest. Moreover, knowledge graphs represent an important scenario to shape and study legal data [30, 92, 54, 31, 43, 23].

Despite such important contributions in visual legal analytics and knowledge graph research, to the best of our knowledge there is still a lack of systems integrating visual and network analysis with recently developed, advanced methods in natural language processing, particularly the deep contextualized language models such as BERT [25]. The latter is instead a key point in the development of our LawNet-Viz.

Chapter 2

Background

In the pursuit of advancing knowledge and addressing complex issues within a specific domain, a comprehensive understanding of the contextual landscape is imperative. The primary objective of this Chapter 2 is to equip the reader with a nuanced understanding of the research domain, fostering a deep appreciation for the intricacies that define the subject.

As the volume of unstructured textual data generated by humanity continues to increase, particularly on the Internet, there is a growing imperative to develop intelligent methods for processing this data and extracting diverse forms of knowledge from it. The primary objective of this doctoral thesis is to advance the field by constructing learning models capable of autonomously eliciting representations of human language, with a specific focus on its structure and meaning. These models aim to address multiple higher-level language tasks through the acquisition of sophisticated language representations.

Significant strides have already been made in the realm of Natural Language Processing (NLP), leading to the advent of technologies proficient in extracting information from vast unstructured datasets on the web, conducting sentiment analysis in social networks, and performing grammatical analysis for essay grading. Within the scope of NLP, a key objective is to pioneer the development of universal and scalable algorithms that can collectively tackle these diverse language-related tasks. The ultimate aim is to cultivate algorithms that can adeptly learn the requisite intermediate representations of linguistic units involved in these tasks.

Despite the significant progress, NLP faces persistent challenges. Ambiguity, context dependency, and the need for large annotated datasets pose ongoing hurdles. Additionally, ethical concerns related to bias in language models and issues of interpretability require careful consideration in the development and deployment of NLP systems.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a multidisciplinary field that combines computer science, artificial intelligence, and linguistics to enable machines to understand, interpret, and generate human language. The primary objective of NLP is to bridge the gap between human communication and computer understanding, allowing machines to interact with users in a manner that is both meaningful and contextually appropriate. Over the years, NLP has witnessed substantial growth, evolving from rule-based approaches to sophisticated deep learning techniques, and has found applications in various domains, including critical ones like legal and health domains. NLP encompasses a diverse range of tasks.

Tokenization. Tokenization is the process of breaking down a text into smaller units, typically words or phrases, referred to as tokens. This foundational task is essential for subsequent NLP tasks, as it establishes the basic units for analysis and understanding of language structure.

Part-of-Speech Tagging. Part-of-speech tagging involves assigning grammatical categories (e.g., noun, verb, adjective) to each token in a sentence. This task is crucial for understanding the syntactic structure of a sentence and is fundamental to many higher-level NLP applications.

Named Entity Recognition (NER). NER focuses on identifying and classifying entities, such as people, locations, organizations, and more, within a given text. This task is vital for information extraction and plays a key role in applications like question-answering systems and knowledge graph construction.

Sentiment Analysis. Sentiment analysis aims to determine the sentiment expressed in a piece of text, whether it is positive, negative, or neutral. This task is valuable in understanding user opinions, customer feedback, and social media content, with applications in business intelligence and market research.

Machine Translation. Machine Translation involves automatically translating text from one language to another. This task has seen significant advancements with the rise of neural machine translation models, which have revolutionized cross-language communication and globalization.

Text Summarization. Text summarization focuses on condensing large volumes of text while retaining the essential information. This task is crucial for handling information overload and has applications in news aggregation, data compression, but it is even essential for subsequent NLP tasks.

Question Answering. Question Answering systems aim to automatically generate relevant answers to user queries based on a given dataset or knowledge base. This task involves understanding the meaning of questions, retrieving relevant information, and generating coherent responses.

Coreference Resolution. Coreference resolution addresses the challenge of identifying when different expressions in a text refer to the same entity. This task is essential for maintaining coherence and clarity in natural language understanding, particularly in discourse and long-form texts.

Entailment. Textual entailment involves determining whether one piece of text logically entails another. This task is crucial for tasks like recognizing inference relationships, and it has applications in information retrieval, question answering, and automated reasoning systems.

2.2 Deep Learning and Transformers in NLP

The advent of statistical and machine learning models in the 1990s marked a significant turning point in NLP. Researchers began leveraging large datasets to train models, enabling systems to automatically learn patterns and relationships within language. Notable achievements during this era include the introduction of Hidden Markov Models (HMMs) for part-of-speech tagging and statistical machine translation models.

The 21st century witnessed the rise of neural network-based approaches, particularly with the introduction of deep learning techniques. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) became pivotal in capturing intricate linguistic structures and dependencies, facilitating advancements in tasks such as sentiment analysis and named entity recognition.

In recent years, the Transformer architecture, introduced by Vaswani et al. [100], has revolutionized NLP by demonstrating remarkable performance in tasks requiring understanding of contextual information. Transformers, relying on self-attention mechanisms, excel in capturing long-range dependencies in sequences, making them well-suited for a wide range of NLP tasks. The introduction of models like BERT [25] and GPT [83] further pushed the boundaries of language understanding and generation.

Transformers Architectures. Natural Language Processing (NLP) has witnessed significant advancements in recent years, with the emergence of transformer-based models playing a pivotal role. Transformers, introduced by Vaswani et al. [100], have revolutionized the field by capturing contextual relationships in sequential data efficiently. These models are

particularly well-suited for NLP tasks due to their ability to handle long-range dependencies and capture intricate patterns in language.

The transformer architecture is built on the mechanism of self-attention, allowing the model to weigh different parts of the input sequence differently. This attention mechanism enables transformers to excel in tasks such as language modeling, translation, and sentiment analysis. The transformer architecture has become the foundation for numerous state-of-the-art models in NLP.

One of the most notable applications of transformers in NLP is the Bidirectional Encoder Representations from Transformers (BERT) model [25]. BERT, a pre-trained transformer model, achieves remarkable results in various downstream tasks by learning contextualized representations of words. The pre-training strategy of BERT involves predicting missing words in a sentence, enabling the model to grasp nuanced language structures.

Subsequent transformer-based models, such as GPT-3 [16] and T5 [84], have further pushed the boundaries of NLP capabilities. GPT-3, for instance, is a language model with an unprecedented number of parameters, allowing it to perform diverse tasks with minimal task-specific fine-tuning. T5, on the other hand, introduces the concept of pre-training on a variety of tasks and fine-tuning for specific tasks, showcasing the versatility of transformer-based architectures.

In addition to their success in natural language understanding, transformers have also demonstrated effectiveness in natural language generation tasks. Multi-modal models like OpenAI's GPT-4 [77] exhibit remarkable language generation capabilities, producing coherent and contextually relevant text. The continuous evolution of transformer-based models in NLP highlights their adaptability and robustness across various tasks.

The core of the Transformer architecture consists of an encoder-decoder structure. The encoder processes input sequences, while the decoder generates output sequences. Each encoder and decoder layer is composed of two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network.

The self-attention mechanism enables each element in the input sequence to focus on different parts of the sequence, capturing long-range dependencies. In the multi-head self-attention, the input sequence is linearly transformed into multiple representations, each attending to different parts of the sequence. The attention weights are computed using a scaled dot-product attention mechanism, providing a way to assign different importance levels to different elements in the sequence.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

After attention, a position-wise feed-forward network is applied independently to each position. This network consists of two fully connected layers with a ReLU activation in between. The position-wise nature allows the network to capture local patterns in the data.

The training of Transformers involves optimizing a task-specific objective function using techniques like stochastic gradient descent (SGD) or Adam optimization. The model is trained to minimize a loss function, typically cross-entropy for classification tasks. Additionally, Transformers often use regularization techniques such as dropout to prevent overfitting. There are several types of Transformers-based architectures.

BERT. BERT, or Bidirectional Encoder Representations from Transformers, a groundbreaking model in natural language processing, introduced a paradigm shift by leveraging bidirectional context understanding. This advanced approach allowed BERT to consider both preceding and succeeding words in a given sequence, enabling a deeper and more comprehensive comprehension of contextual relationships. At the core of BERT's success lies its innovative pre-training strategies, primarily the masked language modeling and the next sentence prediction objectives.

The masked language modeling objective is a distinctive feature of BERT's pre-training phase. During training, a certain percentage of the input tokens in a sequence are randomly selected and replaced with a special [MASK] token. The model is then tasked with predicting the original identity of the masked tokens based on the context provided by the surrounding words. This approach encourages BERT to develop a nuanced understanding of word relationships, capturing not only the immediate context but also the broader syntactic and semantic connections within the sequence. By training on both left and right context simultaneously, BERT excels at capturing bidirectional dependencies, allowing it to excel in tasks requiring a deep understanding of context.

Next, the next sentence prediction (NSP) modeling is another ingenious aspect of BERT's pre-training strategy. BERT is trained on pairs of sentences, and during this process, it learns to predict whether the second sentence in a pair follows the first in the original document. This task encourages the model to grasp the relationships between sentences, promoting a more profound understanding of discourse structure. While this objective was designed for document-level understanding, it indirectly contributes to capturing contextual nuances within individual sentences. By incorporating this task, BERT not only excels in word-level context comprehension but also demonstrates a superior ability to comprehend relationships between sentences, making it well-suited for a wide array of natural language understanding tasks.

The bidirectional context understanding achieved through the masked language modeling and next sentence prediction objectives has been pivotal in BERT's success across a spectrum of NLP benchmarks. This detailed pre-training strategy empowers BERT to create contextually rich representations of language, fostering improvements in various downstream tasks, including question answering, named entity recognition, sentiment analysis, and more. The nuanced bidirectional understanding cultivated by BERT's pre-training has set a standard for subsequent transformer-based models and has significantly contributed to the advancement of natural language processing capabilities.

GPT. The Generative Pre-trained Transformer (GPT) introduced a unique approach to language modeling known as autoregressive language modeling. This method stands in contrast to the bidirectional context understanding employed by models like BERT. In autoregressive language modeling, the model generates text in a sequential manner, predicting the next word in a sequence based on the context of the preceding words. This sequential generation allows GPT to capture dependencies and context in a left-to-right fashion, creating coherent and contextually relevant text.

At the heart of GPT's architecture is the decoder-only transformer. Unlike the original transformer model, which utilizes both encoder and decoder components, GPT relies solely on the decoder architecture for language modeling. The decoder is designed to generate text autoregressively, one word at a time, by attending to the previously generated tokens in the sequence. This sequential generation process aligns well with the nature of language and enables GPT to create contextually rich and coherent outputs.

The decoder-only architecture in GPT is comprised of multiple layers of self-attention mechanisms, allowing the model to weigh the importance of different parts of the input sequence while generating each token. This mechanism enables GPT to capture long-range dependencies in the text, fostering a deeper understanding of context. The self-attention mechanism also contributes to GPT's ability to generate diverse and contextually appropriate completions for a given prompt.

GPT's autoregressive language modeling, coupled with the decoder-only architecture, has proven highly effective in various natural language understanding and generation tasks. The model's proficiency in generating human-like text stems from its ability to learn intricate patterns and structures in sequential data. This architecture has influenced subsequent models and approaches in the field, demonstrating the viability and power of autoregressive language modeling for capturing context in a sequential manner. GPT's contributions have significantly advanced the capabilities of generative models in natural language processing, setting new benchmarks for language generation tasks.

T5. The Text-To-Text Transfer Transformer (T5) represents a significant leap in the field of natural language processing by introducing a unified and versatile framework for various NLP tasks. At the core of T5's innovation is its groundbreaking approach of formulating all NLP tasks as a text-to-text problem. Unlike previous models that specialized in specific tasks like classification or generation, T5 casts every task as a text transformation, where the input and output are both treated as sequences of text. This innovative paradigm shift provides a unified and consistent methodology, simplifying the architecture and training of models for a wide range of tasks.

In the T5 framework, the model is trained to map an input text to an output text, regardless of the nature of the task. Whether it is translation, summarization, question answering, or any other NLP task, T5 learns to represent the task as a text pair where the input represents the source text, and the output represents the target text. This universal representation enables T5 to seamlessly adapt to different tasks without task-specific modifications to the model architecture.

The text-to-text approach offers several advantages. It unifies the pre-training and fine-tuning phases, streamlining the training process. It also allows for straightforward application of the model to new tasks without the need for task-specific architectures or training procedures. The simplicity and consistency brought about by the text-to-text paradigm make T5 an elegant solution for addressing a multitude of NLP challenges.

The versatility of T5 is further enhanced by the ability to easily incorporate new tasks into the training pipeline. Tasks are treated as prompts, and the model is trained to generate the target output based on the given prompt. This flexibility enables researchers and practitioners to extend T5's capabilities to novel tasks without extensive re-engineering, making it a highly adaptable and future-proof model.

T5's text-to-text transfer learning paradigm has had a profound impact on the natural language processing community, influencing subsequent models and approaches. It not only simplifies the development and deployment of models across diverse tasks but also fosters a deeper understanding of the commonalities and shared representations underlying various NLP challenges. The success of T5 underscores the potential benefits of unifying NLP tasks under a common framework, opening avenues for more efficient and generalized language models.

BART. The Bidirectional and Auto-Regressive Transformers (BART) stands as a pioneering model in natural language processing, introducing a novel pre-training objective that combines denoising autoencoders with a bidirectional encoder-decoder architecture. This unique combination of components has significantly contributed to BART's success in various NLP tasks, including abstractive summarization and text generation.

The first key component of BART's training methodology is the denoising autoencoder. In this approach, the model is trained to corrupt and then reconstruct documents. During the pre-training phase, a random subset of the input tokens is masked, similar to the masked language modeling objective in BERT. The model is then tasked with reconstructing the original document by predicting the masked tokens. This denoising autoencoder objective encourages BART to learn a robust and generalized representation of language by focusing on the ability to fill in missing or corrupted information. Unlike traditional autoencoders, BART's denoising autoencoder operates bidirectionally, enhancing its capacity to capture contextual dependencies in both directions.

The second pivotal component of BART's training methodology is its bidirectional encoder-decoder architecture. While the encoder-decoder architecture is commonly associated with sequence-to-sequence tasks like translation, BART employs it during pre-training as well. The bidirectional encoder processes the corrupted input document, capturing contextual information, and the decoder generates the reconstructed document. This bidirectional approach ensures that the model learns to capture both global and local context, fostering a deep understanding of relationships within the text. The encoder-decoder architecture inherently introduces a form of bidirectionality, allowing BART to excel in tasks requiring a nuanced grasp of context, coherence, and semantic meaning.

The bidirectional encoder-decoder architecture, coupled with the denoising autoencoder objective, empowers BART to generate high-quality and coherent text. The model's ability to learn bidirectional context during pre-training sets the stage for its strong performance in downstream tasks, where understanding dependencies within and across sentences is crucial. This innovative combination of components has positioned BART as a versatile and powerful model for a range of NLP applications.

BART's success in abstractive summarization, document completion, and other text generation tasks underscores the effectiveness of its novel pre-training methodology. By seamlessly integrating denoising autoencoders with a bidirectional encoder-decoder architecture, BART has made substantial contributions to advancing the state-of-the-art in natural language processing, demonstrating the potential of combining diverse training objectives for creating more capable language models.

2.3 LLMs: Large Language Models

Developing a comprehensive understanding of human language on a large scale is a challenging and resource-demanding undertaking. The journey to achieve the present capabilities of language models and expansive linguistic frameworks has unfolded over numerous decades.

With the continuous expansion of model sizes, their intricacy and effectiveness have grown proportionally. In the early stages, language models could forecast the probability of a single word. In contrast, contemporary large language models exhibit the ability to predict the probability of entire sentences, paragraphs, or even entire documents.

The dimensions and functionalities of language models have undergone a remarkable surge in recent years, propelled by advancements in computer memory, dataset scale, processing capabilities, and the refinement of techniques for modeling extended sequences of text.

The definition is vague,but “large” has been used to describe BERT (110M parameters) as well as PaLM 2 [6] (up to 340B parameters). The term "large" refers to the quantity of parameters within the model, where the parameters are the weights the model learned during training, employed for predicting the subsequent token in the sequence. LLMs are based on the Transformers architecture and are trained using a pre-training and fine-tuning approach.

ChatGPT. ChatGPT represents a state-of-the-art development in the field of conversational AI. ChatGPT, developed by OpenAI, is a language model based on the GPT-4 architecture. It is a part of the GPT (Generative Pre-trained Transformer) series, known for its ability to generate coherent and contextually relevant text based on input prompts. ChatGPT is specifically tailored for handling conversational contexts, making it a valuable asset in various NLP applications, ranging from chatbots to language understanding tasks. The GPT-3.5 is one of the largest (in terms of number of parameters) model created until now wit 175 billion parameters and the GPT-4 extends this size.

2.3.1 Prompt Engineering.

Prompt engineering is the art of asking the right question to get the best output from an LLM. It enables direct interaction with the LLM using only natural language prompts.

In the past, working with machine learning models typically required deep knowledge of datasets and programming language. Today, LLMs can be “used” in English, as well as other languages. There are several types of prompts.

Zero-shot. The most basic prompt type is zero-shot. It involves presenting the model with only the instruction, without any accompanying examples. Alternatively, you can frame the instruction as a question or assign a “role” to guide the model with instructions and contextual information.

One-, few-, and multi-shot. Prompting with few- or multi-shots involves presenting the model with additional examples of the desired task. This approach is more effective than

zero-shot when dealing with complex tasks requiring pattern replication or when the output needs to adhere to a specific, challenging-to-describe structure.

Chain-of-thought prompting. Chain of Thought (CoT) prompting encourages the model to articulate its thought process. When paired with few-shot prompting, this combination yields improved results for intricate tasks that demand reasoning prior to generating a response.

2.4 Data

In this section we talk about the two textual corpus used for all the research work proposed. They are the Italian Civil Code, currently in force in the Italian civil legal system, and the General Data Protection Regulation, the most important law about data protection in Europe which affects all the world.

2.4.1 ICC: The Italian Civil Code

The Italian Civil Code, hereinafter referred to as ICC, is the legislation source containing norms that regulate private law in Italy and it consists of 2969 articles. This number actually corresponds to 3225 articles considering all variants and subsequent insertions, which are designated by using Latin-term suffixes (e.g., “bis”, “ter”, “quater”). Then, we obtain a total of 3039 if we remove the articles no longer in force (i.e., articles which are replaced by other laws). Enacted by Royal decree no. 262 of March 16, 1942, the ICC has been involved in a perpetual process of refinements and enhancements, and subjected to numerous reorganizations to stay updated with respect to legislative needs and social development. Indeed, during its history, the ICC was revised several times and subjected to repealings, i.e., per-article partial or total insertions, modifications and removals; to date, 2294 articles have been repealed. The ICC is compiled as an organic corpus relating to the fundamental and constitutional civil laws. In addition to the organization into various books and their sections, the corpus and its constituent articles have an additional structure that is described by *cross-article references*. These are citations that may occur in the content of an article to refer to one or more articles, from either the same or different books, and hence they are exploited by legislators to clarify the scope and the semantics of specific articles. The Italian Civil Code (ICC) is divided into six, logically coherent books, each in charge of providing rules for a particular civil law theme:

- *Book-1*, on Persons and the Family, articles 1-455 – contains the discipline of the juridical capacity of persons, of the rights of the personality, of collective organizations, of the family;

- *Book-2*, on Successions, articles 456-809 – contains the discipline of succession due to death and the donation contract;
- *Book-3*, on Property, articles 810-1172 – contains the discipline of ownership and other real rights;
- *Book-4*, on Obligations, articles 1173-2059 – contains the discipline of obligations and their sources, that is mainly of contracts and illicit facts (the so-called civil liability);
- *Book-5*, on Labor, articles 2060-2642 – contains the discipline of the company in general, of subordinate and self-employed work, of profit-making companies and of competition;
- *Book-6*, on the Protection of Rights, articles 2643-2969 – contains the discipline of the transcription, of the proofs, of the debtor's financial liability and of the causes of pre-emption, of the prescription.

The articles of each book are internally organized into a hierarchical structure based on four levels of division, namely (from top to bottom in the hierarchy): "titoli" (i.e., chapters), "capi" (i.e., subchapters), "sezioni" (i.e., sections), and "paragrafi" (i.e., paragraphs). It should however be emphasized that this hierarchical classification was not meant as a crisp, ground-truth organization of the articles' contents: indeed, the topical boundaries of contiguous chapters and subchapters are often quite smooth, as articles in the same group often not only vary in length but can also provide dispositions that are more related to articles in other groups.

The ICC is obviously publicly available, in various digital formats. From one of such sources, we extracted *article id*, *title* and *content* of each article. We cleaned up the text from non-ASCII characters, removed numbers and date, normalized all variants and abbreviations of frequent keywords such as "articolo" (i.e., article), "decreto legislativo" (i.e., legislative decree), "Gazzetta Ufficiale" (i.e., Official Gazette), and finally we lowercased all letters.

Table 2.1 summarizes main statistics on the preprocessed ICC books.

2.4.2 GDPR: General Data Protection Regulation

The General Data Protection Regulation (GDPR) stands as one of the most significant legal frameworks for data protection and privacy in recent years. Enforced by the European Union (EU) since May 2018, the GDPR has garnered global attention due to its wide-reaching impact on businesses, organizations, and individuals, transcending geographical boundaries.

Table 2.1 Main statistics on the ICC corpus and its constituent books.

ICC portion	# arts.	# sentences over the articles				# words over the articles			
		tot.	*min*	*max*	*mean* (*std*)	*tot.*	*min*	*max*	*mean* (*std*)
Book-1	395	1979	3	21	5.010 (2.323)	32354	11	569	81.909 (71.952)
Book-2	345	1561	3	13	4.525 (1.675)	24520	9	354	71.072 (51.366)
Book-3	364	1619	3	24	4.448 (1.816)	25971	6	893	71.349 (65.836)
Book-4	891	3595	3	12	4.035 (1.338)	50509	7	365	56.688 (38.837)
Book-5	713	3937	3	37	5.522 (3.191)	75764	8	1465	106.261 (117.393)
Book-6	331	1453	3	17	4.390 (1.895)	25937	12	654	78.360 (76.954)
All	3039	14131	3	37	4.650 (2.243)	234945	6	1465	77.310 (78.373)

While initially conceived to safeguard the data rights of EU citizens, its influence extends far beyond EU member states, making it a pivotal legislation worldwide.

The GDPR consists of two components, namely the *articles* and *recitals*, whose main characteristics are described next.

Articles. The GDPR articles constitute the legal requirements that must be followed by organizations to demonstrate compliance. The GDPR currently in force consists of 99 articles, which are organized into 11 chapters:

• *Chapter I*, on General provisions, articles 1-4 – sets out the overall aims and scope of the GDPR, including its territorial scope and the definitions of key terms used throughout the regulation;

• *Chapter II*, on Principles, articles 5-11 – lays out the fundamental principles that organizations must follow when collecting, processing, and storing personal data, including lawfulness, fairness, transparency, and purpose limitation;

• *Chapter III*, on Rights of the data subject, articles 12-23 – outlines the rights of individuals with regards to their personal data, including the right of access, the right to rectification, the right to erasure, the right to restriction of processing, and the right to data portability;

• *Chapter IV*, on Controller and Processor, articles 24-43 – sets out the obligations of data controllers and processors, including the obligation to appoint a data protection officer and to conduct data protection impact assessments;

• *Chapter V*, on Transfers of personal data to third countries or international organisations, articles 44-50 – sets out the rules for transferring personal data to third countries or international organizations, including the use of standard contractual clauses and binding corporate rules;

• *Chapter VI*, on Independent supervisory authorities, articles 51-59 – establishes the role of independent supervisory authorities and sets out their powers, including the power to investigate, impose administrative fines, and order the suspension of data processing;

• *Chapter VII*, on Cooperation and Consistency, articles 60-76 – sets out the obligations

of data controllers and processors to cooperate with supervisory authorities and to ensure consistent application of the GDPR;
• *Chapter VIII*, on Remedies, liability and penalties, articles 77-84 – sets out the remedies available to individuals whose rights have been violated, including the right to receive compensation, and sets out the potential sanctions for organizations that breach the GDPR, including administrative fines and criminal sanctions;
• *Chapter IX*, on Provisions relating to specific processing situations, articles 85-91 – sets out specific rules for certain types of processing, including processing for scientific or historical research purposes, processing for the performance of a task carried out in the public interest, and processing for the establishment, exercise, or defense of legal claims;
• *Chapter X*, on Delegated acts and implementing acts, articles 92-93 – sets out the powers of the European Commission to adopt delegated acts and implementing acts to further specify certain aspects of the GDPR;
• *Chapter XI*, on Final provisions, articles 94-99 – sets out the final provisions of the GDPR, including its entry into force.

Recitals. In addition to the articles, GDPR is comprised of a set of introductory statements and explanations, called *recitals*, that provide context and guidance for the interpretation of the provisions of the regulation. They indeed provide valuable insights into the underlying objectives and policy considerations, and represent a valuable resource for determining the meaning and scope of the GDPR articles. The recitals are 173 in total, each associated to one or more articles; in turn, each article is associated with zero, one or more recitals.

An example of the importance of recitals for GDPR articles can be found in the interpretation of the article 7(2) 'Conditions for consent', which emphasizes the importance of clear and transparent communication when obtaining consent for the processing of personal data. A guidance on how controllers can meet this requirement under the GDPR can be clarified and expanded upon by the recital 42, which is related to the article itself. This recital states that, when processing is based on the data subject's consent, the controller must be able to demonstrate that the data subject has given consent to the processing operation. This means that the controller must have records or other evidence to prove that valid consent was obtained; furthermore, to be considered informed, the data subject must be aware of the identity of the controller and the purposes of the processing for which their personal data is intended.

The GDPR hence encompasses a comprehensive set of norms and principles that regulate the collection, processing, and transfer of personal data. Its provisions, such as the right to be forgotten, consent requirements, and data subject rights, have brought about significant changes in the digital landscape. Indeed, the complexity and scope of the GDPR pose

challenges for government agencies, law firms, legal professionals, and citizens seeking to navigate its intricacies and access relevant regulations.

Chapter 3

LamBERTa: Law article mining based on BERT architecture applied on the ICC

This chapter is based on the research work presented in the papers "Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code" [97] , "LamBERTa: Law Article Mining Based on Bert Architecture for the Italian Civil Code" [96] and "Exploring domain and task adaptation of LamBERTa models for article retrieval on the Italian Civil Code" [90]. Firstly, in Section 3.1, we delve into a comprehensive exploration of the novel LamBERTa framework. Following this, we transition into a rigorous Experimental Evaluation, in Section 3.2. Section 3.3 unfolds as we present the Results derived from our experiments, shedding light on the outcomes and insights gleaned from the application of the LamBERTa Framework. Building upon these findings, we study the effects coming from the combination of domain adaptive pre-training strategies and task adaptive training strategies in Section 3.4. Furthermore, in Section 3.5 we scrutinize the concept of Explainability leveraging on the use of different techniques.

3.1 The Proposed LamBERTa Framework

In this section we present our proposed learning framework for civil-law article retrieval. We first formulate the problem statement in Section 3.1.1 and overview the framework in Section 3.1.2. We present our devised learning approaches in Section 3.1.3. We describe the data preparation and preprocessing in Section 3.1.4, then in Section 3.1.5 we define our unsupervised training-instance labeling methods for the articles in the target corpus. Finally, in Section 3.1.6, we discuss major settings of the proposed framework.

3.1.1 Problem setting

Our study is concerned with law article retrieval, i.e., finding articles of interest out of a legal corpus that can be recommended as an appropriate response to a query expressing a legal matter.

To formalize this problem, we assume that any query is expressed in natural language and discusses a legal subject that is in principle covered by the target legal corpus (i.e., the ICC, in our context). Moreover, a query is assumed to be free of references to any article identifier in the ICC.

We address the law article retrieval task based on the supervised machine learning paradigm: given a new, user-provided instance, i.e., a legal question, the goal is to automatically predict the category associated to the posed question. More precisely, we deal with the more general case in which a probability distribution over all the predefined categories is computed in response to a query. The prediction is carried out by a machine learning system that is trained on a target legal corpus — whose documents, i.e., articles, are annotated with the actual category they belong to — in order to learn a computational model, also called classifier, that will then be used to perform the predictions against legal queries by exclusively utilizing the textual information contained in the annotated documents.

Motivations for BERT-based approach. In this respect, our objective is to leverage deep neural-network-based, pre-trained language modeling to solve the law article retrieval task. This has a number of key advantages that are summarized as follows. First, like any other deep neural network models, it totally avoids manual feature engineering, and hence the need for employing feature selection methods as well as feature relevance measures (e.g., TF-IDF). Second, like sophisticated recurrent and convolutional neural networks, it models language semantics and non-linear relationships between terms; however, better than recurrent and convolutional neural networks and their combinations, it is able to capture subtle and complex lexical patterns including the sequential structure and long-term dependencies, thus obtaining the most comprehensive local and global feature representations of a text sequence. Third, it incorporates the so-called *attention* mechanism, which allows a learning model to assign higher weight to text features according to their higher informativeness or relevance to the learning task. Fourth, being a real bidirectional *Transformer* model, it overcomes the main limitations of early deep contextualized models like ELMO [78] (whose left-to-right language model and right-to-left language model are actually independently trained) or decoder-based Transformer models, like OpenAI GPT [83].

Challenges. It should however be noted that, like any other machine learning method, using deep pre-trained models like BERT for classification tasks normally requires the availability

of data annotated with the class labels, so to design the independent training and testing phases for the classifier. However, this does not apply to our context.

As previously mentioned, in this work we face three main challenges:

- the first challenge refers to the high number (i.e., hundreds) of classes, which correspond to the number of articles in the ICC corpus, or portion of it, that is used to train a LamBERTa model;
- the second challenge corresponds to the so-called *few-shot learning* problem, i.e., dealing with a small amount of per-class examples to train a machine learning model, which Bengio et al. recognize as one of the "extreme classification" scenarios [9].
- the third challenge derives from the unavailability of test query benchmarks for Italian legal article retrieval/prediction tasks. This has prompted us to define appropriate methods for data annotation, thus for building up training sets for the LamBERTa framework. To address this problem, we originally define different schemes of *unsupervised training-instance labeling*; notably, these are not ad-hoc defined for the ICC corpus, rather they can be adapted to any other law code corpus.

In the following, we overview our proposed deep pre-trained model based framework that is designed to address all the above challenges.

3.1.2 Overview of the LamBERTa framework

Figure 3.1 shows the conceptual architecture of our proposed LamBERTa – **L**aw **a**rticle **m**ining based on **BERT a**rchitecture. The starting point is a pre-trained Italian BERT model whose source data consists of a recent Wikipedia dump, various texts from the OPUS corpora collection, and the Italian part of the OSCAR corpus; the final training corpus has a size of 81GB and 13 138 379 147 tokens.[1]

LamBERTa models are generated by fine-tuning the pre-trained Italian BERT model on a sequence classification task (i.e., BERT with a single linear classification layer on top) given in input the articles of the ICC or a portion of it.

It is worth noting that the LamBERTa architecture is versatile w.r.t. the adopted learning approach and the training-instance labeling scheme for a given corpus of ICC articles. These aspects are elaborated on next.

[1] bert-base-italian-xxl-uncased, available at https://huggingface.co/dbmdz/.

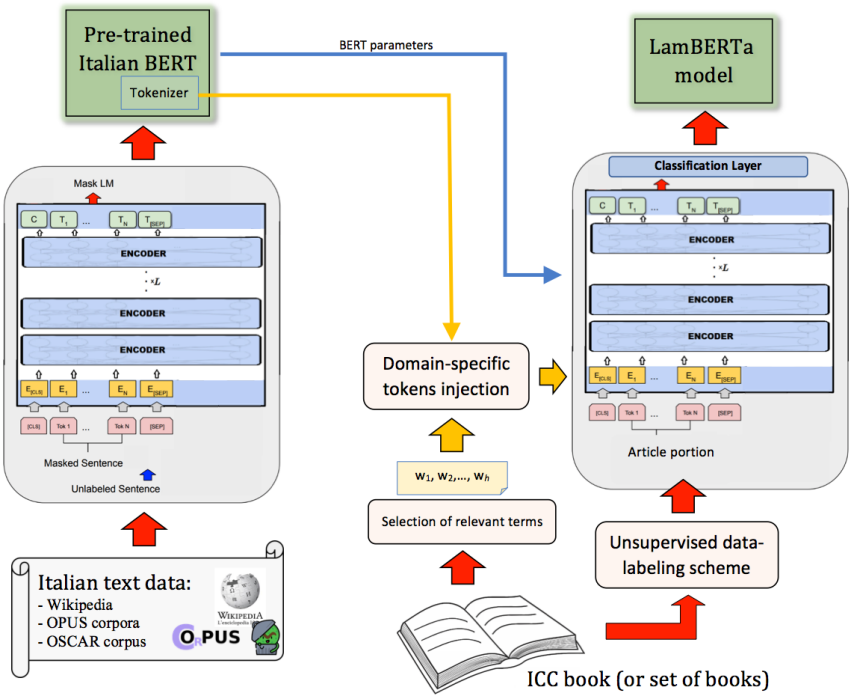


Fig. 3.1 An illustration of the conceptual architecture of LamBERTa designed for the ICC law article retrieval task.

3.1.3 Global and Local learning approaches

We consider two learning approaches, here dubbed *global* and *local* learning, respectively. A **global model** is trained on the whole ICC, whereas a **local model** is trained on a particular book of the ICC.[2] Our rationale underlying this choice is as follows:

- on the one hand, local models are designed to embed the logical coherence of the articles within a particular book and, although limited to its corresponding topical boundaries, they are expected to leverage the multi-faceted semantics underlying a specific civil law theme (e.g., inheritage);
- on the other hand, books are themselves part of the same law code, and hence a global model might be useful to capture possible interrelations between the single books, however, by embedding different topic signals from different books (e.g. inheritage of Book-2 vs. labor law of Book-5), it could incur the risk of topical dilution over all the ICC.

[2]From a technical viewpoint, the considered portion could in principle include any strict subset of the ICC; however, on the legal domain side, it is reasonable to learn local models each dedicated to a logically coherent subset of the whole civil code, i.e., a single book.

Either type of model is designed to be a *classifier at article level*, i.e., class labels correspond to the articles in the book(s) covered by the model. Given the one-to-one association between classes and articles, a question becomes how to create suitable training sets for our LamBERTa models. Our key idea is to adopt an *unsupervised annotation* approach, which is discussed in the next section.

3.1.4 Data preparation

To tailor the ICC articles to the BERT input format, we initially carried out segmentation of the content of each article into sentences. This was then followed by tokenization of the sentences and text encoding, which are described next.

Domain-specific terms injection and Tokenization. BERT was trained using the WordPiece tokenization. This is an effective way to alleviate the open vocabulary problems in neural machine translation, since a word can be broken down into sub-word units, which are constructed during the training time and depend on the corpus the model was initially trained on.

However, when retraining BERT on the ICC articles to learn global as well as local models, it is likely that some important terms occurring in the legal texts are missing in the pre-trained lexicon; therefore, to make BERT aware of the domain-specific (i.e., legal) terms and avoid subwording such terms thus disrupting their semantics, we injected a selection of terms from the ICC articles into the pre-existing BERT vocabulary before tokenization.[3]

To select the domain-specific terms to be added, we carried out the following steps: if we denote with D the input (portion of) ICC text for the model to learn, we first preprocessed the text D as described in Sect. 2.4, then we removed Italian stopwords, and finally filtered out overly frequent terms (as occurring in more than 50% of the articles in D) as well as hapax terms. Table 3.1 reports the number of added tokens and the final number of tokens, for each input corpus to our LamBERTa local and global models.

Text encoding. BERT utilizes a fixed sequence size (usually 512) for each tokenized text, which implies both padding of shorter sequences and truncation of longer sequences. While padding has no side effect, truncation may produce loss of information.

In learning our models, we wanted to avoid the above aspect, therefore we investigated any condition causing such undesired effect in our input data. We found out that this contingency is very rare in each book of the ICC, even reducing the maximum length to 256 tokens; the only situation that would lead to truncation corresponds to an article sentence that is logically

[3] From the perspective of the BERT architecture, this implies two actions, namely indexing the added terms and initializing their corresponding weights.

Table 3.1 Number of domain-specific tokens to inject into the pre-trained BERT vocabulary and the final vocabulary size.

ICC portion	# added tokens	vocab. size
Book-1	833	31935
Book-2	698	31800
Book-3	1072	32174
Book-4	1383	32485
Book-5	2048	33150
Book-6	829	31931
All	3993	35095

organized in multiple clauses separated by semicolon: for these cases, we treated each of the subsentences as one training unit associated with the same article class-label.

3.1.5 Methods for unsupervised training-instance labeling

As previously discussed, our LamBERTa classification models are trained in such a way that it holds an one-to-one correspondence between articles in a target corpus and class labels. Moreover, the entire corpus must be used to train the model so to embed the whole knowledge therein. However, given the uniqueness of each article, the general problem we face is *how to create as many training instances as possible for each article* to effectively train the models.

Our key idea is to select and combine portions of each article to generate the training units for it. To this aim, we devise different unsupervised schemes for data labeling the ICC articles to create the training sets of LamBERTa models. Our schemes adopt different strategies for selecting and combining portions of each article to derive the training sets, but they share the requirements of generating a minimum number of training units per article, here denoted as $minTU$; moreover, since each article is usually comprised of few sentences, and $minTU$ needs to be relatively large (we chose 32 as default value), each of the schemes implements a *round-robin* (RR) method that iterates over replicas of the same group of training units per article until at least $minTU$ are generated.

In the following, we define our methods for unsupervised training-instance labeling of the ICC articles (in square brackets, we indicate the notation that will be used throughout the remainder of the paper):

- *Title-only* [T]. This is the simplest yet lossy scheme, which keeps an article's title while discarding its content; the round-robin block is just the title of an article.[4]

- *n-gram*. Each training unit corresponds to n consecutive sentences of an article; the round-robin block starts with the n-gram containing the title and ends with the n-gram containing the last sentence of the article. We set $n \in \{1,2,3\}$, i.e., we consider a unigram [UniRR], a bigram [BiRR], and a trigram [TriRR] model, respectively.

- *Cascade* [CasRR]. The article's sentences are cumulatively selected to form the training units; the round-robin block starts with the first sentence (i.e., the title), then the first two sentences, and so on until all article's sentences are considered to form a single training unit.

- *Triangle* [TglRR]. Each training unit is either an unigram, a bigram or a trigram, i.e., the round-robin block contains all n-grams, with $n \in \{1,2,3\}$, that can be extracted from the article's title and description.

- *Unigram with parameterized emphasis on the title* [UniRR.T$^+$]. The set of training units is comprised of one subset containing the article's sentences with round-robin selection, and another subset containing only replicas of the article's title. More specifically, the two subsets are formed as follows:

  - The first subset is of size equal to the maximum between the number of article's sentences and the quantity $m \times mean\_s$, where m is a multiplier (set to 4 as default) and *mean_s* expresses the average number of sentences per article, excluding the title. As reported in Table 2.1 (sixth column), this mean value can be recognized between 3 and 4 – recall that the title is excluded from the count – therefore we set $mean\_s \in \{3,4\}$.
  - The second subset finally contains $minTU - m \times mean\_s$ replicas of the title.

- *Cascade with parameterized emphasis on the title* [CasRR.T$^+$] and *Triangle with parameterized emphasis on the title* [TglRR.T$^+$]. These two schemes follow the same approach as UniRR.T$^+$except for the composition of the round-robin block, which corresponds to CasRR and TglRR, respectively, with the title left out from this block and replicated in the second block, for each article.

[4]We point out that considering also the title of the group to which an article belongs (e.g., chapter, subchapter) would not add further information, since each article's title recalls yet specializes its group's title.

3.1.6 Learning configuration

An input to LamBERTa is of the form [CLS, ⟨*text*⟩, SEP], where CLS (stands for 'classification') and SEP are special tokens respectively at the beginning of each sequence and at separation of two parts of the input. These two special tokens are associated with two vectorial dense representations or *embeddings*, denoted as $E_{[CLS]}$ and $E_{[SEP]}$, respectively; analogously, there are as many embeddings E_i as the number of tokens of the input text. Note also that as discussed in Sect. 3.1.4, each sequence in a mini-batch is padded to the maximum length (e.g., 256 tokens) in the batch. In correspondence of each input embedding E_i, an embedding T_i is outputted, and it can be seen as the contextual representation of the i-th token of the input text. Instead, the final hidden state, i.e., output embedding C, corresponding to the CLS token captures the high level representation of the entire text to be used as input to further levels in order to train any sentence-based task, such as sentence classification. This vector, which is by design of size 768, is fed to a single layer neural network followed by sigmoid activation whose output represents the distribution probability over the law articles of a book.

LamBERTa models were trained using a typical configuration of BERT for masked language modeling, with 12 attention heads and 12 hidden layers, and initial (i.e., pre-trained) vocabulary of 32 102 tokens. Each model was trained for 10 epochs, using cross-entropy as loss function, Adam optimizer and initial learning rate selected within [1e-5, 5e-5] on batches of 256 examples.

It should be noted that our LamBERTa models can be seen as relatively light in terms of computational requirements, since we developed them under the Google Colab GPU-based environment with a limited memory of 12 GB and 12-hour limit for continuous assignment of the virtual machine.[5]

3.2 Experimental Evaluation

In this section we present the methodology and results of an experimental evaluation that we have thoroughly carried out on LamBERTa models. In the following, we first state our evaluation goals in Section 3.2.1, then we define the types of test queries and select the datasets in Section 3.2.2, finally we describe our methodology and assessment criteria in Section 3.2.3.

[5] https://colab.research.google.com/.

3.2.1 Evaluation goals

Our main evaluation goals can be summarized as follows:

- To validate and measure the effectiveness of LamBERTa models for law article retrieval tasks: how do local and global models perform on different evaluation contexts, i.e., against queries of different type, different length, and different lexicon? (Sect. 3.3.1)
- To evaluate LamBERTa models in single-label as well as multi-label classification tasks: how do they perform w.r.t. different assumptions on the article relevance to a query, particularly depending on whether a query is originally associated with or derived from a particular article, or by-definition associated with a group of articles? (Sect. 3.3.1)
- To understand how a LamBERTa model's behavior is affected by varying and changing its constituents in terms of training-instance labeling schemes and learning parameters (Sect. 3.3.2).
- To demonstrate the superiority of our classification-based approach to law article retrieval by comparing LamBERTa to other deep-learning-based text classifiers (Sect. 3.3.3) and to a few-shot learner conceived for an attribute-aware prediction task that we have newly designed based on the ICC heading metadata (Sect. 3.3.3).

3.2.2 Query sets

A query set is here meant as a collection of natural language texts that discuss legal subjects relating to the ICC articles. For evaluation purposes, each query as a test instance is associated with one or more article-class labels.

It should be emphasized that, on the one hand, we cannot devise a training-test split or cross-validation of the target ICC corpus (since we want that our LamBERTa models embed the knowledge from all articles therein), and on the other hand, there is a lack of query evaluation benchmarks for Italian civil law documents. Therefore, we devise different types of query sets, which are aimed at representing testbeds at varying difficulty level for evaluating our LamBERTa models:

- QType-1 – *book-sentence-queries* refer to a set of queries that correspond to randomly selected sentences from the articles of a book. Each query is derived from a single article, and multiple queries are from the same article.

- QType-2 – *paraphrased-sentence-queries* share the same composition of QType-1 queries but differ from them as the sentences of a book's articles are paraphrased. To this purpose, we adopt a simple approach based on backtranslation from English (i.e., an original sentence in Italian is first translated to English, then the obtained English sentence is translated to Italian).[6]

- QType-3 – *comment-queries* are defined to leverage the publicly available *comments* on the ICC articles provided by legal experts through the platform "Law for Everyone".[7] Such comments are delivered to provide annotations about the interpretation of the meanings and law implications associated to an article, or to particular terms occurring in an article. Each query corresponds to a comment available about one article, which is a paragraph comprised of about 5 sentences on average.

- QType-4 – *comment-sentence-queries* refer to the same source as QType-3, but the comments are split into sentences, so that each query contains a single sentence of a comment. Therefore, each query will be associated to a single article, and multiple queries will refer to the same article.

- QType-5 – *case-queries* refer to a collection of case law decisions from the civil section of the Italian Court of Cassation, which is the highest court in the Italian judicial system. These case law decisions are selected from publicly available corpora of the most significant jurisprudential sentences associated with the ICC articles, spanning over the period 1977-2015.

- QType-6 – *ICC-heading-queries* are defined by extracting the headings of chapters, subchapters, and sections of each ICC book. Such headings are very short, ranging from one to few keywords used to describe the topic of a particular division of a book.

Characteristics and differences of the query-sets. It should be noted that while QType-1 and QType-6 query sets have the same lexicon as the corresponding book articles, this does not necessarily hold for QType-2 due to the paraphrasing process, and the difference becomes more evident with QType-3, QType-4, and QType-5 query sets. Indeed, the latter not only originate from a corpus different from the ICC, but they have the typical verbosity of annotations or comments about law articles as well as of case law decisions; moreover, they often provide cross-book references, therefore, QType-3, QType-4, and QType-5 query-set contents are not necessarily bounded by a book's context.

[6] We used the Google Translate service, which is a widely-used yet effective machine translator.

[7] *Law for Everyone* online news portal, available at https://www.laleggepertutti.it.

Table 3.2 Main statistics on the test query sets.

ICC portion	query type	#queries	#words	query type	# queries	#words	#sentences
Book-1	QType-1	790	13 436	QType-3	331	38 383	1 262
Book-2	QType-2	690	10 998	QType-4	322	30 226	1 000
Book-3		728	11 942		286	23 815	781
Book-4		1 774	27 352		764	85 360	2 640
Book-5		1 426	22 754		570	87 094	2 417
Book-6		662	11 993		294	31 898	1 040

ICC portion	query type	#queries	#words	year range
Book-1	QType-5	333	39 464	1978-2014
Book-2		347	36 856	1979-2014
Book-3		371	39 643	1979-2014
Book-4		975	111 234	1978-2015
Book-5		1 037	114 287	1977-2015
Book-6		720	81 186	1978-2015

ICC portion	query type	#chapter queries	#subchapter queries	#section queries	#total words
Book-1	QType-6	14	25	22	317
Book-2		5	30	14	160
Book-3		9	21	29	230
Book-4		9	51	57	366
Book-5		11	11	51	397
Book-6		5	18	38	251

Moreover, QType-3, QType-5, and QType-6 differ from the other types in terms of length of each query, which corresponds indeed to multiple sentences or a paragraph, in the case of QType-3 and QType-5, and to a single or few keywords, in the case of QType-6. Note that, although derived from the ICC books, *the contents of the QType-6 queries were totally discarded when training our LamBERTa models*; moreover, unlike the other types of queries, each QType-6 query is by definition associated with a group of articles (i.e., according to the book divisions) rather than a single article, therefore we shall use QType-6 queries for the multi-label evaluation task only. Also, it should be emphasized that the QType-5 queries represent a different difficult testbed as they contain real-life, heterogeneous fact descriptions of the case and judicial precedents.

Table 3.2 summarizes main statistics on the query sets. In addition, note that the percentage of QType-1 that were paraphrased (i.e., to produce QType-2 queries) resulted in above 85% for each of the books. Also, the number of sentences in a book's query set (last column of the upper subtable) corresponds to the number of queries of QType-4 for that book.

It is also worth emphasizing that all query sets were validated by legal experts. This is important to ensure not only generic linguistic requirements but also meaningfulness of the query contents from a legal viewpoint.

3.2.3 Evaluation methodology and Assessment criteria

Let us denote with $\mathscr{C}$ either a portion (i.e., a single book) of the ICC, or the whole ICC. We specify an evaluation context in terms of a test set (i.e., query set) Q pertinent to $\mathscr{C}$ and a LamBERTa model $\mathscr{M}$ learnt from $\mathscr{C}$. Note that the pertinence of Q w.r.t. $\mathscr{C}$ is differently determined depending on the type of query set, as previously discussed in Section 3.2.2.

We consider two multi-class evaluation contexts: *single-label* and *multi-label*. In the single-label context, each query is pertinent to only one article, therefore there is only one relevant result for each query. In the multi-label context, each query can be pertinent to more than one article, therefore there can be multiple relevant results for each query.

Single-label evaluation context

We consider a confusion matrix in which the article ids correspond to the classes. Upon this matrix, we measure the following standard statistics for each article $A_i \in \mathscr{C}$: the precision for A_i (P_i), i.e., the number of times (queries) A_i was correctly predicted out of all predictions of A_i, the recall for A_i (R_i), i.e., the number of times (queries) A_i was correctly predicted out of all queries actually pertinent to A_i, the F-measure for A_i (F_i), i.e., $F_i = 2P_iR_i/(P_i+R_i)$. Then, we averaged over all articles to obtain the per-article average *precision* (P), *recall* (R), and two types of F-measure: *micro-averaged F-measure* (F^{μ}) as the actual average over all F_is, and *macro-averaged F-measure* (F^M) as the harmonic mean of P and R, i.e., $F^M = 2PR/(P+R)$.

In addition, we consider two other types of criteria that respectively account for the top-k predictions and the position (rank) of the correct article in predictions. The first aspect is captured in terms of the fraction of correct article labels that are found in the top-k predictions (i.e., top-k-probability results in response to each query), and averaging over all queries, which is the *recall@k* ($R@k$). The second aspect is measured as the *mean reciprocal rank* (*MRR*) considering for each query the rank of the correct prediction over the classification probability distribution, and averaging over all queries.

Moreover, we wanted to understand the *uncertainty of prediction* of LamBERTa models when evaluated over each query: to this purpose, we measured the *entropy* of the classification probability distributions obtained for each query evaluation, and averaged over all queries;

in particular, we distinguished between the entropy of each entire distribution (E) from the entropy of the distribution corresponding to the top-k-probability results ($E@k$).

Multi-label evaluation context

The basic requirement for the multi-label evaluation context is that, for each test query, a set of articles is regarded as relevant to the query. In this respect, we consider two different perspectives, depending on whether a query is (i) originally associated with or derived from a particular article, or (ii) by-definition associated with a group of articles. We will refer to the former scenario as *article-driven* multi-label evaluation, and to the second scenario as *topic-driven* multi-label evaluation.

Article-driven multi-label evaluation. For this stage of evaluation, we require that, for each test query associated with article A_i, a set of articles *related* to A_i is to be selected as the set of articles relevant to the query, together with A_i. We adopt two approaches to determine the *article relatedness*, which rely on a supervised and an unsupervised grouping of the articles, respectively. The supervised approach refers to the logical organization of the articles of each book originally available in the ICC (cf. Section 2.4), whereas the unsupervised approach leverages a content-similarity-based grouping of the articles produced by a document clustering method.

ICC-classification-based. We exploit the ICC-provided classification meta-data of each book to label each article with the terminal division it belongs to. Since a division usually contains a few articles, the same label will be associated to multiple articles. It should be noted that the book hierarchies are quite different to each other, both in terms of maximum branching (e.g., from 5 chapters in Book-2 and Book-6, to 14 chapters in Book-1) and total number of divisions (e.g., from 51 in Book-2 to 135 in Book-4); moreover, sections, subchapters or even chapters can be leaf nodes in the corresponding hierarchy, i.e., they contain articles without any further division. This leads to the following distribution of number of labels over the various books: 50 in Book-1, 40 in Book-2, 47 in Book-3, 112 in Book-4, 94 in Book-5, 56 in Book-6.

Clustering-based. As concerns the unsupervised, similarity-based approach, we compute a clustering of the set of articles of a given book, so that each cluster will correspond to a group of articles that are similar by content. Each of the produced clusters will be regarded as the set of relevant articles for each query whose original label article belongs to that set. Therefore, if a cluster contains n articles, then each of the queries originally labeled with any of such n articles, will share the same set of n relevant articles.

To perform the clustering of the article set of a given book, we resort to a widely-used, well-known document clustering method, which consists in applying a centroid-based

partitional clustering algorithm [41] over a document-term matrix modeling a vectorial bag-of-words representation of the documents over the term feature space, using TF-IDF (term-frequency inverse-document-frequency) as term relevance weighting function [42] and cosine similarity for document comparison. This method is also known as *spherical k-means* [26]. We used a particularly effective and optimized version of this method, called bisecting k-means [114], which is a standard de-facto of document clustering tools based on the classic bag-of-words model.[8]

It should be emphasized that our goal here is to induce an organization of the articles that is based on content affinity while discarding any information on the ICC article labeling. In this regard, we have deliberately exploited a representation model that, despite its known limitations in being unable to capture latent semantic aspects underlying correlations between words, it is still an effective baseline, yet unbiased with respect to the deep language model ability of LamBERTa.

Computation of relevant sets. Let us denote with $\mathscr{P}$ a partitioning of $\mathscr{C}$ that corresponds to either the ICC classification of the articles in $\mathscr{C}$ (i.e., supervised organization) or a clustering of $\mathscr{C}$ (i.e., unsupervised organization). For either approach, given a test query $q_i \in Q$ with article label A, we detect the relevant article set for q as the partition $P \in \mathscr{P}$ that contains A, then we match this set to the set of top-$|P|$ predictions of $\mathscr{M}$ to compute precision, recall, and F-measure for q_i. Finally, we averaged over all queries to obtain overall precision, recall, micro-averaged F-measure and macro-averaged F-measure.

Topic-driven multi-label evaluation. Unlike the previously discussed evaluation stage, here we consider queries that are expressed by one or few keywords describing the topic associated with a set of articles.

To this purpose, we again exploit the meta-data of article organization available in the ICC books, so that the articles belonging to the same division of a book (i.e., chapter, subchapter, section) will be regarded as the set of articles relevant to the query corresponding to the description of that division. Depending on the choice of the type of a book's division, i.e., the level of the ICC-classification hierarchy of that book, a different query set will be produced for the book.

Given a test query $q_i \in Q$ with article labels $A_{i_1}, \ldots, A_{i_n}$, we match this set to the set of top-i_n predictions of $\mathscr{M}$ to compute precision, recall, and F-measure for q_i. Finally, we averaged over all queries to obtain overall precision, recall, micro-averaged F-measure and macro-averaged F-measure. Moreover, we measured the fraction of top-k predictions that are relevant to a query, and averaging over all queries, i.e., *precision*@k ($P@k$).

[8] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview.

3.3 Results

We organize the presentation of results into three parts. Section 3.3.1 describes the comparison between global and local models under both the single-label and multi-label evaluation contexts; note that we will use notations G and L_i to refer to the global model and the local model corresponding to the i-th book, respectively, for any given test query-set. Section 3.3.2 is devoted to an ablation study of our models, focusing on the analysis of the various unsupervised data labeling methods and the effect of the training size on the models' performance. Finally, Section 3.3.3 presents results obtained by competing deep-learning methods.

3.3.1 Global vs. Local models

Single-label evaluation

Table 3.3 and Fig. 3.2 compare global and local models, for all books of the ICC, according to the criteria selected for the single-label evaluation context. For the sake of brevity yet representativeness, here we show results corresponding to the UniRR.T$^+$ labeling scheme — as we shall discuss in Section 3.3.2, this choice of training-instance labeling scheme is justified as being in general the best-performing scheme for the query sets; nonetheless, analogous findings were drawn by using other types of queries and labeling schemes.

Looking at the table, let us first consider the results obtained by the LamBERTa models against QType-1 queries. At a first sight, we observe the outstanding scores achieved by all models; this was obviously expected since QType-1 queries are directly extracted from sentences of the book articles used to train the models. More interesting is the comparison between the local models and the global model: for any query-set from the i-th book, the corresponding local model behaves significantly better than the global model in most cases. It should be noticed that the slightly better results of the global model in terms of precision are actually determined by its occasional wrongly predictions of articles from different books in place of articles of Book-i.

The outperformance of local models over the global one is also confirmed in terms of the entropy results, which are always lower, and hence better, than the global's ones. Moreover, while the plots in Fig. 3.2 are representative for QType-1 and QType-4 queries, the above result for the comparison between local and global models in terms of entropy behavior equally holds regardless of the type of query set. Note that, in Fig. 3.2(a) and (d), the values of entropy are normalized within the interval [0, 1] to get a fair comparison between the local

Table 3.3 Global vs. local models, for all sets of book-sentence-queries (QType-1, upper subtable), paraphrased-sentence-queries (QType-2, second upper subtable), comment-queries (QType-3, third upper subtable), comment-sentence-queries (QType-4, second bottom subtable), and case-queries (QType-5, bottom subtable): Recall, Precision, micro- and macro-averaged F-measures, Recall@k, and MRR. (Bold values correspond to the best model for each query-set and evaluation criterion)

	R		P		F^{μ}		F^{M}		R@3		R@10		MRR	
i	L_i	G	L_i	G	L_i	G	L_i	G	L_i	G	L_i	G	L_i	G
Q_1	**0.962**	0.949	0.975	**0.979**	**0.961**	0.955	**0.969**	0.964	0.989	**0.992**	**0.999**	0.997	**0.976**	0.970
Q_2	**0.972**	0.954	0.981	**0.985**	**0.971**	0.961	**0.977**	0.969	**0.994**	0.981	**0.999**	0.986	**0.983**	0.968
Q_3	**0.990**	0.978	**0.994**	0.994	**0.990**	0.981	**0.992**	0.986	**1.000**	0.995	**1.000**	0.997	**0.995**	0.987
Q_4	**0.961**	0.948	0.983	**0.985**	**0.964**	0.958	**0.972**	0.966	**0.984**	0.977	**0.990**	0.986	**0.975**	0.966
Q_5	**0.910**	0.905	0.944	**0.947**	**0.909**	0.908	**0.927**	0.925	**0.984**	0.974	**0.998**	0.988	**0.947**	0.940
Q_6	**0.979**	0.961	**0.986**	0.986	**0.978**	0.966	**0.982**	0.973	**1.000**	0.985	**1.000**	0.992	**0.989**	0.974
Q_1	**0.841**	0.730	**0.866**	0.823	**0.828**	0.745	**0.853**	0.774	**0.905**	0.829	**0.949**	0.881	**0.881**	0.784
Q_2	**0.828**	0.639	**0.856**	0.766	**0.814**	0.669	**0.841**	0.697	**0.992**	0.742	**0.941**	0.812	**0.871**	0.701
Q_3	**0.861**	0.728	**0.886**	0.843	**0.851**	0.756	**0.873**	0.781	**0.922**	0.816	**0.942**	0.842	**0.896**	0.778
Q_4	**0.736**	0.635	**0.756**	0.706	**0.713**	0.639	**0.746**	0.668	**0.806**	0.716	**0.861**	0.783	**0.779**	0.685
Q_5	**0.718**	0.684	**0.759**	0.742	**0.710**	0.686	**0.738**	0.712	**0.843**	0.808	**0.908**	0.867	**0.790**	0.755
Q_6	**0.841**	0.704	**0.874**	0.817	**0.833**	0.730	**0.857**	0.756	**0.914**	0.783	**0.941**	0.845	**0.882**	0.756
Q_1	**0.349**	0.25	**0.248**	0.196	**0.274**	0.211	**0.290**	0.220	**0.494**	0.422	**0.675**	0.530	**0.455**	0.355
Q_2	**0.313**	0.177	**0.213**	0.145	**0.239**	0.155	**0.253**	0.159	**0.494**	0.307	**0.655**	0.441	**0.445**	0.265
Q_3	**0.396**	0.260	**0.298**	0.216	**0.325**	0.228	**0.340**	0.236	**0.577**	0.426	**0.704**	0.574	**0.507**	0.371
Q_4	**0.336**	0.247	**0.239**	0.190	**0.264**	0.206	**0.279**	0.215	**0.487**	0.387	**0.622**	0.522	**0.438**	0.343
Q_5	0.171	**0.252**	0.128	**0.190**	0.137	**0.205**	0.147	**0.216**	0.364	**0.433**	0.562	**0.590**	0.302	**0.375**
Q_6	**0.387**	0.235	**0.291**	0.178	**0.317**	0.193	**0.332**	0.203	**0.588**	0.388	**0.745**	0.551	**0.515**	0.340
Q_1	**0.190**	0.136	**0.197**	0.183	**0.170**	0.135	**0.194**	0.156	**0.363**	0.265	**0.502**	0.376	**0.333**	0.239
Q_2	**0.216**	0.100	**0.191**	0.148	**0.173**	0.106	**0.203**	0.119	**0.343**	0.172	**0.473**	0.257	**0.315**	0.157
Q_3	**0.241**	0.153	**0.223**	0.188	**0.204**	0.145	**0.231**	0.169	**0.433**	0.268	**0.569**	0.405	**0.395**	0.250
Q_4	**0.189**	0.144	**0.176**	0.173	**0.156**	0.137	**0.182**	0.157	**0.324**	0.253	**0.450**	0.345	**0.293**	0.228
Q_5	0.092	**0.119**	0.132	**0.148**	0.090	**0.113**	0.108	**0.132**	0.256	**0.278**	0.414	**0.426**	0.222	**0.248**
Q_6	**0.224**	0.144	**0.211**	0.209	**0.192**	0.147	**0.218**	0.171	**0.372**	0.237	**0.508**	0.336	**0.333**	0.210
Q_1	**0.228**	0.118	**0.233**	0.121	**0.210**	0.101	**0.230**	0.120	**0.494**	0.304	**0.813**	0.488	**0.430**	0.263
Q_2	**0.292**	0.162	**0.316**	0.232	**0.274**	0.172	**0.303**	0.191	**0.501**	0.280	**0.726**	0.444	**0.465**	0.270
Q_3	**0.284**	0.190	**0.323**	0.217	**0.273**	0.180	**0.302**	0.203	**0.566**	0.330	**0.856**	0.474	**0.489**	0.286
Q_4	**0.259**	0.159	**0.299**	0.187	**0.250**	0.150	**0.278**	0.172	**0.528**	0.314	**0.803**	0.493	**0.458**	0.279
Q_5	**0.354**	0.256	**0.401**	0.307	**0.342**	0.252	**0.376**	0.279	**0.641**	0.484	**0.884**	0.647	**0.564**	0.426
Q_6	**0.392**	0.241	**0.445**	0.313	**0.372**	0.247	**0.417**	0.273	**0.665**	0.376	**0.885**	0.520	**0.590**	0.347

models and global models.[9] This indicates that the predictions made by a local model are generally more certain than those by the global model.

Turning back to the results reported in Table 3.3, we observe that high performance scores are still obtained for the paraphrased-sentence-based queries, which indicates a remarkable robustness of LamBERTa models w.r.t. lexical variants of queries.

[9]For each book, the entropy value of the local model is divided by the base-2 logarithm of the number of articles of that book, whereas the entropy value of the global model is always divided by the base-2 logarithm of the total number of articles in the ICC.

Considering now the comment-based queries, both in the paragraph-size (i.e., QType-3) and sentence-size versions (i.e., QType-4), we observe a much lower performance of the models. This is nonetheless not surprising since QType-3 and QType-4 queries are lexically and linguistically different from the ICC contents (cf. Section 3.2.2). Besides that, local models show to be consistently better than the global ones, despite the cross-corpora learning ability of the global model which, in principle, could be beneficial for queries that may contain references to articles of different books. One exception corresponds to Book-5: as confirmed by a qualitative inspection of the legal experts who assisted us, this is due to an accentuation in the queries for Book-5 of a tendency in more broadly covering subjects of articles from other books. This is also supported by quantitative statistics about the uniqueness of the terms in the vocabulary of a book: in fact, we found out that the percentage of a book's vocabulary terms that are unique is minimum w.r.t. Book-5; in detail, 39% of Book-1 and Book-2, 48% of Book-3, 43% of Book-4, and 36% of Book-6. A further interesting remark can be drawn from the comparison of the QType-3 scores with the QType-4 scores, as these appear to be generally higher for the former queries, which provide more topical context than sentence-size queries. This would hence suggest that relatively long queries can be handled by LamBERTa models effectively.

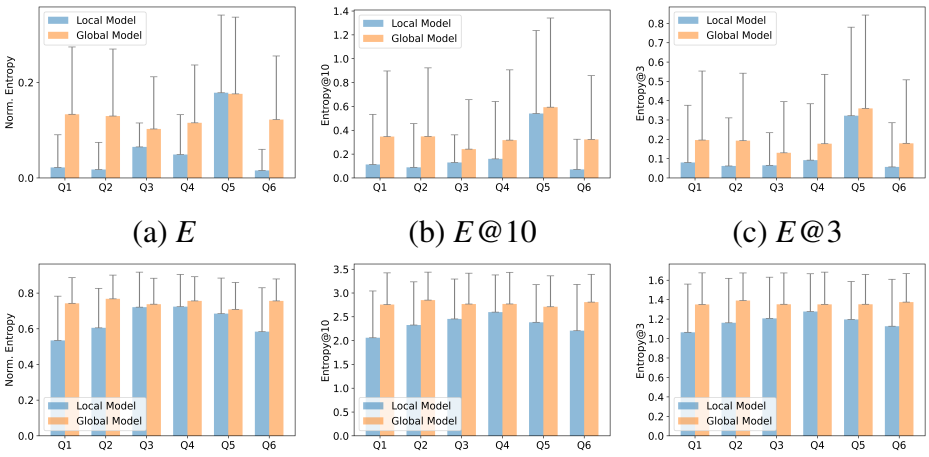
Notably, the above finding on the ability of handling long, contextualized queries is also confirmed by inspectioning the results obtained against the queries stating case law decisions (i.e., QType-5 queries). Indeed, despite the particular difficulty of such a testbed, LamBERTa models show performance scores that are generally comparable to those obtained for the QType-3 comment queries; moreover, once again, local models behave better than the corresponding global ones, according to all assessment criteria (and even without exceptions as opposed to QType-3 comment queries).

Multi-label evaluation

Our investigation on the effectiveness of global and local models was further deepened under the different multi-label evaluation contexts we devised.

Article-driven multi-label evaluation. Table 3.4 and Table 3.5 report results according to the article-driven clustering-based and ICC-classification-based analyses, respectively. For the clustering-based approach, results correspond to a number of clusters over the articles of each book that was set so to have mean cluster size close to three.

At a first glance, the superiority of local models stands out, thus confirming the findings drawn from the single-label evaluation results. Upon a focused inspection, we observe few exceptions in Table 3.4 corresponding to results on Book-5 (similarly as we already found in Table 3.3) for the QType-1, QType-3, and QType-4, and on Book-4 for QType-2; however,



(a) E (b) $E@10$ (c) $E@3$

(d) E (e) $E@10$ (f) $E@3$

Fig. 3.2 Global vs. local models, for all sets of book-sentence-queries (a-c) and comment-sentence-queries (d-f): Mean and standard deviation of the normalized Entropy and of the Entropy$@k$ values of the classification probability distributions

this does not occur in Table 3.5. We tend to ascribe this to the fact that, being driven by a content-similarity-based approach, the clustering of the articles is clearly conditioned on the higher topic variety observable in Book-5 or Book-4, which may favor a global model against a local one in better capturing topics that are outside the boundaries of that particular book. By contrast, this does not hold for the ICC-classification-based grouping of the articles, as it relies on an externally provided organization of the articles in a particular book that is not constrained by the topic patterns (e.g., word co-occurrences) that might be distinctive of that book.

Notably, looking at the QType-5 results, the performance scores obtained by local and global models are generally comparable to or even higher than those respectively obtained on QType-3 (or QType-4) queries, which is particularly evident for the ICC-classification-based grouping of the articles (Table 3.5).

As a further point of investigation, we explored whether the article-driven clustering-based multi-label evaluation is sensitive to how the query relevant sets (i.e., clusters of articles) were formed, with a focus on the content representation model of the articles. In particular, we replaced the TF-IDF vectorial representation of the articles with the article embeddings generated by our LamBERTa models, while keeping the same clustering methodology and setting as used in our previous analysis. From the comparison results reported in the Appendix, Table 3.15, it stands out that using the embeddings generated by LamBERTa models to represent the articles in the clustering process is always beneficial to the multi-label classification performance of LamBERTa local models, according to all criteria and query

Table 3.4 Article-driven, Clustering-based multi-label evaluation: global vs. local models, with labeling scheme UniRR.T$^+$, for all sets of book-sentence-queries (QType-1), paraphrased-sentence-queries (QType-2), comment-queries (QType-3), comment-sentence-queries (QType-4), and case-queries (QType-5). (Bold values correspond to the best model for each query-set and evaluation criterion)

query-set		R		P		F^{μ}		F^{M}	
	i	L_i	G	L_i	G	L_i	G	L_i	G
QType-1	Q_1	**0.615**	0.592	**0.617**	0.591	**0.615**	0.591	**0.616**	0.591
	Q_2	**0.674**	0.636	**0.684**	0.640	**0.676**	0.637	**0.679**	0.638
	Q_3	**0.623**	0.604	**0.626**	0.603	**0.624**	0.603	**0.625**	0.603
	Q_4	**0.331**	0.313	**0.326**	0.312	**0.324**	0.312	**0.328**	0.312
	Q_5	0.631	**0.632**	0.633	**0.634**	0.628	**0.631**	0.632	**0.633**
	Q_6	**0.711**	0.675	**0.720**	0.677	**0.710**	0.675	**0.715**	0.676
QType-2	Q_1	**0.557**	0.473	**0.559**	0.474	**0.557**	0.472	**0.558**	0.473
	Q_2	**0.595**	0.442	**0.604**	0.445	**0.598**	0.443	**0.599**	0.444
	Q_3	**0.563**	0.477	**0.566**	0.479	**0.563**	0.477	**0.564**	0.478
	Q_4	0.172	**0.236**	0.213	**0.237**	0.186	**0.236**	0.190	**0.237**
	Q_5	**0.522**	0.498	**0.526**	0.499	**0.522**	0.497	**0.524**	0.498
	Q_6	**0.624**	0.511	**0.628**	0.512	**0.625**	0.511	**0.626**	0.513
QType-3	Q_1	**0.312**	0.241	**0.315**	0.242	**0.312**	0.241	**0.313**	0.242
	Q_2	**0.296**	0.164	**0.300**	0.165	**0.297**	0.164	**0.298**	0.164
	Q_3	**0.332**	0.226	**0.337**	0.228	**0.333**	0.226	**0.334**	0.227
	Q_4	**0.234**	0.174	**0.241**	0.175	**0.236**	0.174	**0.237**	0.174
	Q_5	0.209	**0.249**	0.214	**0.252**	0.209	**0.250**	0.211	**0.251**
	Q_6	**0.361**	0.234	**0.365**	0.236	**0.361**	0.234	**0.363**	0.235
QType-4	Q_1	**0.227**	0.154	**0.227**	0.155	**0.226**	0.154	**0.227**	0.154
	Q_2	**0.205**	0.098	**0.207**	0.103	**0.205**	0.096	**0.206**	0.100
	Q_3	**0.259**	0.153	**0.261**	0.154	**0.258**	0.153	**0.260**	0.153
	Q_4	**0.163**	0.123	**0.164**	0.127	**0.164**	0.123	**0.163**	0.125
	Q_5	0.151	**0.172**	0.152	**0.179**	0.150	**0.174**	0.151	**0.175**
	Q_6	**0.236**	0.147	**0.238**	0.149	**0.235**	0.147	**0.237**	0.148
QType-5	Q_1	**0.283**	0.179	**0.289**	0.183	**0.284**	0.180	**0.286**	0.181
	Q_2	**0.294**	0.170	**0.304**	0.175	**0.298**	0.172	**0.299**	0.173
	Q_3	**0.298**	0.171	**0.302**	0.174	**0.299**	0.171	**0.300**	0.172
	Q_4	**0.274**	0.192	**0.276**	0.193	**0.275**	0.191	**0.275**	0.192
	Q_5	**0.378**	0.296	**0.382**	0.298	**0.378**	0.295	**0.380**	0.297
	Q_6	**0.404**	0.226	**0.408**	0.229	**0.405**	0.226	**0.406**	0.227

sets. As expected, the improvements are generally more evident for QType-1 and QType-2 query sets, since these queries have lexicons close to the training data. More interestingly, we also observe that the performance gain by LamBERTa embeddings tends to be higher for the largest books, i.e., Book-4 and Book-5, with peaks of about 300% percentage increase reached with the QType-2 query-set relevant for Book-4. Nonetheless, apart from these particular cases, discovering clusters over a set of articles represented by their LamBERTa embeddings does not bring in general to outstanding boost of performance against the TF-IDF vectorial representation: indeed, this can be ascribed to the very small size of the clusters

Table 3.5 Article-driven, ICC-classification-based multi-label evaluation: global vs. local models, with labeling scheme UniRR.T$^+$, for all sets of book-sentence-queries (QType-1), paraphrased-sentence-queries (QType-2), comment-queries (QType-3), comment-sentence-queries (QType-4), and case-queries (QType-5). (Bold values correspond to the best model for each query-set and evaluation criterion)

query-set		R		P		F^{μ}		F^{M}	
	i	L_i	G	L_i	G	L_i	G	L_i	G
QType-1	Q_1	**0.247**	0.220	**0.312**	0.269	**0.268**	0.236	**0.276**	0.242
	Q_2	**0.198**	0.187	**0.256**	0.237	**0.216**	0.203	**0.223**	0.209
	Q_3	**0.217**	0.192	**0.287**	0.250	**0.238**	0.209	**0.247**	0.217
	Q_4	**0.191**	0.179	**0.234**	0.217	**0.205**	0.191	**0.210**	0.196
	Q_5	**0.201**	0.191	**0.253**	0.240	**0.218**	0.207	**0.224**	0.213
	Q_6	**0.268**	0.243	**0.306**	0.283	**0.279**	0.254	**0.286**	0.262
QType-2	Q_1	**0.244**	0.194	**0.310**	0.239	**0.265**	0.209	**0.273**	0.214
	Q_2	**0.215**	0.160	**0.278**	0.200	**0.235**	0.173	**0.246**	0.178
	Q_3	**0.206**	0.164	**0.280**	0.224	**0.229**	0.182	**0.238**	0.189
	Q_4	**0.172**	0.144	**0.213**	0.178	**0.186**	0.156	**0.190**	0.159
	Q_5	**0.186**	0.168	**0.235**	0.213	**0.202**	0.183	**0.208**	0.188
	Q_6	**0.257**	0.212	**0.297**	0.248	**0.268**	0.222	**0.276**	0.228
QType-3	Q_1	**0.245**	0.167	**0.319**	0.215	**0.269**	0.183	**0.277**	0.188
	Q_2	**0.170**	0.096	**0.239**	0.134	**0.192**	0.108	**0.199**	0.112
	Q_3	**0.189**	0.133	**0.265**	0.186	**0.212**	0.149	**0.220**	0.155
	Q_4	**0.169**	0.126	**0.215**	0.161	**0.184**	0.138	**0.189**	0.141
	Q_5	**0.154**	0.146	**0.204**	0.200	**0.169**	0.165	**0.176**	0.169
	Q_6	**0.240**	0.167	**0.285**	0.211	**0.252**	0.178	**0.261**	0.186
QType-4	Q_1	**0.188**	0.122	**0.241**	0.153	**0.205**	0.132	**0.211**	0.136
	Q_2	**0.123**	0.065	**0.174**	0.090	**0.139**	0.073	**0.144**	0.076
	Q_3	**0.156**	0.094	**0.230**	0.133	**0.176**	0.105	**0.184**	0.110
	Q_4	**0.126**	0.094	**0.164**	0.122	**0.139**	0.103	**0.142**	0.106
	Q_5	**0.118**	0.115	**0.188**	0.176	**0.140**	0.134	**0.145**	0.139
	Q_6	**0.173**	0.116	**0.204**	0.143	**0.182**	0.123	**0.187**	0.129
QType-5	Q_1	**0.309**	0.199	**0.381**	0.237	**0.334**	0.212	**0.341**	0.216
	Q_2	**0.193**	0.110	**0.265**	0.151	**0.215**	0.122	**0.223**	0.127
	Q_3	**0.244**	0.167	**0.274**	0.179	**0.255**	0.172	**0.258**	0.173
	Q_4	**0.233**	0.158	**0.296**	0.199	**0.254**	0.172	**0.261**	0.176
	Q_5	**0.214**	0.174	**0.272**	0.224	**0.234**	0.191	**0.240**	0.196
	Q_6	**0.245**	0.159	**0.276**	0.185	**0.253**	0.165	**0.259**	0.171

produced, and hence the TF-IDF vectorial representation can well approximate and match the clusters detected over the LamBERTa embeddings of the articles, especially in smaller collections of articles (i.e., ICC books).

Topic-driven multi-label evaluation. Let us now focus on the topic-driven multi-label evaluation results, based on QType-6 queries, which are shown in Table 3.6 (details on precision and recall values are omitted for the sake of presentation, but this does not change the remarks being discussed). Three major remarks arise here. The first one is again about the higher effectiveness of local models against the global one, for each of the book query-sets,

Table 3.6 Topic-driven multi-label evaluation: global vs. local models, with labeling scheme `UniRR.T`$^{+}$, for all sets of ICC-heading-queries (i.e., QType-6 query sets). Column 'a_cs' stands for average class size, i.e., the average no. of articles belonging to each query label. (Bold values correspond to the best model for each query-set and evaluation criterion)

	chapter					subchapter					section				
i	a_cs	F^{μ}		P@3		a_cs	F^{μ}		P@3		a_cs	F^{μ}		P@3	
		L_i	G	L_i	G		L_i	G	L_i	G		L_i	G	L_i	G
Q_1	26.1	**0.314**	0.251	**0.867**	0.733	14.4	**0.307**	0.253	**0.875**	0.792	7.5	**0.348**	0.205	**0.722**	0.556
Q_2	69.0	**0.256**	0.247	0.800	0.800	11.6	**0.167**	0.145	**0.519**	0.370	11.8	**0.234**	0.089	**0.500**	0.333
Q_3	40.3	**0.287**	0.225	**0.667**	0.556	17.6	**0.185**	0.145	**0.778**	0.389	8.5	**0.270**	0.219	**0.577**	0.500
Q_4	98.4	**0.238**	0.083	**0.444**	0.333	17.1	**0.216**	0.144	**0.673**	0.531	8.3	**0.218**	0.149	**0.653**	0.531
Q_5	64.8	**0.174**	0.163	0.545	**0.727**	22.2	**0.263**	0.237	0.733	0.733	10.9	0.215	**0.219**	**0.558**	0.512
Q_6	66.2	**0.402**	0.198	**1.000**	0.600	21.4	**0.319**	0.202	**0.667**	0.467	6.5	**0.290**	0.248	**0.571**	0.486

with the usual exception corresponding to Book-5, which is more evident as the type of division is finer-grain. The second remark concerns a comparison between the methods' performance by varying the types of book division: the chapter, subchapter and section cases are indeed quite different to each other, which should be noticed through the different values of the average number of articles "covered" by each query label-class; in general, higher values correspond to more abstract divisions. Moreover, the roughly monotone variation of this statistic over all books cannot be coupled with a monotonic variation of the performances: for instance, the highest F-measure values correspond to the division at section level in Book-1, whereas they correspond to the chapter level in Books-2, 3, 4 and 6. Finally, it is worth noticing that the performances significantly increase when precision@3 scores are considered, often reaching very high values: this would suggest that, despite the intrinsic complexity of this evaluation task, the models are able to guarantee a high fraction of top-3 predictions that correspond to the articles that are relevant for each query.

3.3.2 Ablation study

Training-instance labeling schemes

Besides the settings of BERT learning parameters, our LamBERTa models are expected to work differently depending on the choice of training-instance labeling scheme (cf. Section 3.1.5). Understanding how this aspect relates to the performance is fundamental, as it impacts on the complexity of the induced model. In this respect, we compared our defined labeling schemes, using the induced local models of Book-2 (i.e., L_2) as case in point.

Table 3.7 shows single-label evaluation results for all types of query-sets. As expected, we observe that the scheme based on an article's title only is largely the worst-performing method (with the exception of QType-2 where, due to the little impact of paraphrasing on the article

Table 3.7 Evaluation of different local models L_2 for all types of query-sets pertinent to articles of Book-2. (Bold values correspond to the best model for each query-set and evaluation criterion)

query-set	method	R	P	F^{μ}	F^{M}	R@10	MRR	E	E@10
QType-1	T	0.530	0.855	0.623	0.655	0.619	0.564	3.023	1.335
	UniRR	0.675	0.928	0.755	0.782	0.851	0.744	0.343	0.167
	BiRR	0.552	0.775	0.621	0.645	0.745	0.635	0.790	0.342
	TriRR	0.584	0.763	0.637	0.661	0.801	0.674	1.149	0.455
	CasRR	0.555	0.901	0.653	0.687	0.712	0.608	1.422	0.808
	TglRR	0.900	0.955	0.910	0.927	0.980	0.929	0.459	0.281
	UniRR.T^{+}	**0.972**	**0.981**	**0.971**	**0.977**	**0.999**	**0.983**	**0.149**	**0.089**
	CasRR.T^{+}	0.823	0.929	0.843	0.873	0.903	0.853	0.863	0.457
	TglRR.T^{+}	0.919	0.972	0.927	0.945	0.977	0.940	0.492	0.298
QType-2	T	0.410	0.598	0.456	0.487	0.551	0.463	3.892	1.706
	UniRR	0.549	0.744	0.605	0.632	0.754	0.624	**1.310**	**0.585**
	BiRR	0.342	0.470	0.374	0.396	0.581	0.436	2.084	0.899
	TriRR	0.442	0.620	0.494	0.516	0.699	0.544	2.253	0.881
	CasRR	0.383	0.620	0.444	0.473	0.597	0.455	2.289	1.232
	TglRR	0.612	0.722	0.626	0.662	0.823	0.684	2.101	1.107
	UniRR.T^{+}	**0.828**	**0.856**	**0.814**	**0.841**	**0.941**	**0.871**	1.620	0.732
	CasRR.T^{+}	0.625	0.740	0.639	0.677	0.784	0.685	2.221	1.074
	TglRR.T^{+}	0.632	0.729	0.642	0.677	0.854	0.705	2.311	1.174
QType-3	T	0.023	0.008	0.010	0.012	0.171	0.073	6.312	2.816
	UniRR	0.296	0.208	0.230	0.244	0.618	0.425	4.368	2.051
	BiRR	0.212	0.140	0.156	0.169	0.494	0.321	5.185	2.349
	TriRR	0.194	0.113	0.133	0.143	0.478	0.295	5.885	2.523
	CasRR	0.110	0.075	0.082	0.089	0.363	0.199	**4.034**	**1.800**
	TglRR	0.261	0.186	0.203	0.217	0.624	0.394	4.048	1.939
	UniRR.T^{+}	**0.313**	**0.213**	**0.239**	**0.253**	**0.655**	**0.445**	4.938	2.266
	CasRR.T^{+}	0.232	0.157	0.176	0.187	0.525	0.339	4.717	2.172
	TglRR.T^{+}	0.217	0.162	0.173	0.185	0.593	0.347	4.254	2.100
QType-4	T	0.037	0.042	0.028	0.039	0.137	0.076	6.379	2.800
	UniRR	0.175	0.199	0.160	0.186	0.439	0.271	4.000	1.834
	BiRR	0.104	0.125	0.094	0.114	0.309	0.185	4.592	2.065
	TriRR	0.090	0.101	0.077	0.096	0.275	0.163	5.357	2.241
	CasRR	0.056	0.068	0.049	0.061	0.219	0.115	**3.268**	**1.615**
	TglRR	0.154	0.178	0.141	0.165	0.416	0.253	4.174	1.991
	UniRR.T^{+}	**0.216**	0.191	**0.173**	**0.203**	**0.473**	**0.315**	5.077	2.311
	CasRR.T^{+}	0.135	0.161	0.119	0.147	0.347	0.211	4.586	2.130
	TglRR.T^{+}	0.144	**0.202**	0.143	0.168	0.398	0.226	4.148	2.027
QType-5	T	0.036	0.025	0.022	0.029	0.127	0.081	6.570	2.968
	UniRR	0.260	0.288	0.232	0.273	0.666	0.436	**4.446**	**2.074**
	BiRR	0.232	0.236	0.197	0.234	0.548	0.357	4.873	2.198
	TriRR	0.225	0.248	0.204	0.236	0.568	0.380	5.498	2.408
	CasRR	0.082	0.089	0.068	0.086	0.334	0.187	5.429	2.400
	TglRR	0.278	**0.318**	0.258	0.297	0.686	0.434	4.560	2.150
	UniRR.T^{+}	**0.292**	0.316	**0.274**	**0.303**	**0.726**	**0.465**	4.972	2.299
	CasRR.T^{+}	0.238	0.295	0.233	0.263	0.579	0.380	4.822	2.158
	TglRR.T^{+}	0.283	0.316	0.273	0.299	0.700	0.463	4.638	2.137

Table 3.8 Single-label evaluation of local model L_2 `UniRR.T`$^+$with $minTU = 64$, for each type of query-set. Percentage values correspond to the increase/decrease percentage of performance criteria when using $minTU = 64$ compared to $minTU = 32$. (Highlighted in red are the values corresponding to a worsening when using $minTU = 64$.)

query-set	R	P	F^{μ}	F^{M}	R@10	MRR	E	E@10
QType-1	0.971	0.982	0.971	0.977	0.996	0.982	0.090	0.072
	-0.10%	+0.13%	-0.04%	-0.04%	-0.34%	-0.13%	-39.40%	-19.41%
QType-2	0.797	0.825	0.781	0.811	0.928	0.845	1.266	0.657
	-3.73%	-3.67%	-4.09%	-3.62%	-1.43%	-2.97%	-21.85%	-10.30%
QType-3	0.351	0.254	0.280	0.295	0.666	0.481	3.661	1.814
	+12.14%	+19.25%	+17.15%	+16.60%	+1.68%	+8.09%	-25.86%	-19.95%
QType-4	0.221	0.204	0.185	0.212	0.482	0.308	3.889	1.928
	+2.31%	+7.37%	+6.94%	+4.95%	+2.55%	-0.96%	-23.40%	-16.57%
QType-5	0.342	0.352	0.320	0.347	0.740	0.460	3.615	1.872
	+17.13%	+11.33%	+16.67%	+14.53%	+1.93%	-1.08%	-27.27%	-18.58%

Table 3.9 Multi-label evaluation of local model L_2 `UniRR.T`$^+$with $minTU = 64$, for each type of query-set. Percentage values correspond to the increase/decrease percentage of performance criteria when using $minTU = 64$ compared to $minTU = 32$. (Highlighted in red are the values corresponding to a worsening when using $minTU = 64$.)

query-set	Clustering-based				ICC-Classification-based			
	R	P	F^{μ}	F^{M}	R	P	F^{μ}	F^{M}
QType-1	0.658	0.668	0.662	0.663	0.197	0.252	0.215	0.221
	-2.08%	-2.20%	-2.07%	-2.21%	-0.51%	-1.56%	-0.46%	-0.90%
QType-2	0.556	0.565	0.559	0.560	0.198	0.252	0.215	0.222
	-6.55%	-6.46%	-6.52%	-6.51%	-8.04%	-9.34%	-8.45%	-8.61%
QType-3	0.322	0.327	0.324	0.324	0.166	0.228	0.185	0.192
	+8.78%	+9.00%	+9.09%	+8.72%	-2.35%	-4.60%	-3.65%	-3.52%
QType-4	0.202	0.206	0.203	0.204	0.131	0.183	0.148	0.153
	-0.98%	+0.00%	-0.98%	-0.49%	+6.50%	+5.17%	+6.47%	+6.25%
QType-5	0.312	0.327	0.319	0.320	0.181	0.247	0.201	0.209
	+6.04%	+7.72%	+7.12%	+7.09%	-6.22%	-6.79%	-6.51%	-6.28%

titles, it happens that the performance of `T` is slightly better than `CasRR`). More interestingly, lower size of n-gram seems to be beneficial to the effectiveness of the model, especially on QType-3, QType-4, and QType-5 queries, indeed `UniRR` always outperforms both `BiRR` and `TriRR`; also, `UniRR` behaves better than the cascade scheme (`CasRR`) as well, and again the gap is more evident in the comment-based queries. The combination of n-grams of varying size reflected by the `TglRR` scheme leads to a significant increase in the performance over all previously mentioned schemes, for QType-1, QType-2, and QType-5 queries. This would suggest that more sophisticated labeling schemes can lead to higher effectiveness in the learned model. Nonetheless, superior performance is obtained by considering the schemes with emphasis on the title, which all improve upon the corresponding schemes not

emphasizing the title, with UniRR.T^{+} being the best-performing method by far according to all criteria.

It also should be noted that the comment and case law query testbeds confirmed to be more difficult than the other types of queries. Also, QType-3 and QType-5 results are found to be generally higher than those obtained for QType-4 queries, which would indicate that, by providing a more informative context, long (i.e., paragraph-like) queries are better handled by LamBERTa models w.r.t. their shorter (i.e., sentence-like) counterparts.

Training units per article

The setting of the number of training units per article (*minTU*, whose default is 32), which impacts on the size of the training sets, is another model parameter that in principle deserves attention. We investigated this aspect according to both single-label and multi-label evaluation contexts, whose results are reported in Table 3.8 and Table 3.9, respectively, for the local model learned from Book-2; note that, once again, this is for the sake of presentation, as we found out analogous remarks for other books.

Our goal was to understand whether and to what extent doubling the *minTU* value could lead to improve the model performance. In this respect, considering the single-label evaluation case, we observe two different outcomes when setting *minTU* to 64: the one corresponding to QType-1 and QType-2 queries, which unveils a general degradation of the performance, and the other one corresponding to QType-3, QType-4, QType-5 queries, which conversely shows significant improvements according to most criteria. This is remarkable as it would suggest that increasing the number of replicas in building the books' training sets is unnecessary, or even detrimental, for queries lexically close to the book articles; by contrast, it turns out to be useful for improving the classifier performance against more difficult query-testbeds like comment and case queries. Looking at the multi-label evaluation results, we can draw analogous considerations for the clustering-based evaluation approach (since results over QType-4 are only slightly decreased), whereas using less training replicas appears to be preferable for the ICC-classification-based evaluation approach.

3.3.3 Comparative analysis

Text-based law article prediction

We conducted a comparative analysis of LamBERTa models with state-of-the-art text classifiers based on deep learning architectures:

- *BiLSTM* [59, 116], a bidirectional LSTM model as sequence encoder. LSTM models have been widely used in text classification as they can capture contextual information while representing the sentence by fixed size vector. The model exploited in this evaluation utilizes 2 layers of BiLSTM, with 32 hidden units for each BiLSTM layer.

- *TextCNN* [45], a convolutional-neural-network-based model with multiple filter widths for text encoding and classification. Every sentence is represented as a bidimensional tensor of shape (n,d), where n is the sentence length and d is the dimensionality of the word embedding vectors. The *TextCNN* model utilizes three different filter windows of sizes $\{3,4,5\}$, 100 feature maps for the windows' sizes, ReLU activation function and max-pooling.

- *TextRCNN* [49], a bidirectional LSTM with a pooling layer on the last sequence output. *TextRCNN* therefore combines the recurrent neural network and convolutional network to leverage the advantages of the individual models in capturing the text semantics. The model first exploits a recurrent structure to learn word representations for every word in the text, thus capturing the contextual information; afterwards, max-pooling is applied to determine which features are important for the classification task.

- *Seq2Seq-A* [28, 7], a Seq2Seq model with attention mechanism. Seq2Seq models have been widely used in machine translation and document summarization due to their capability to generate new sequences based on observed text data. For text classification, here the *Seq2Seq-A* model utilizes a single layer BiLSTM as encoder with 32 hidden units. This encoder learns the hidden representation for every word in an input sentence, and its final state is then used to learn attention scores for each word in the sentence. After learning the attention weights, the weighted sum of encoder hidden states (hidden states of words) gives the attention output vector. The latter is then concatenated to the hidden representation and passed to a linear layer to produce the final classification.

- *Transformer* model for text classification, which is adapted from the model originally proposed for the task of machine translation in [100]. The key aspect in this model is the use of an attention mechanism to deal with long range dependencies, but without resorting to RNN models. The encoder part of the original Transformer model is used for classification. This encoder is composed of 6 layers, each having two sub-layers, namely a multi-head attention layer, and a 2-layer feed-forward network. Compared to BERT, residual connection, layer normalization, and masking are discarded.

Table 3.10 Competing models trained with the individual ICC books annotated with the UniRR.T^{+} labeling scheme, and tested over all sets of book-sentence-queries (QType-1, upper subtable), paraphrased-sentence-queries (QType-2, second upper subtable), comment-sentence-queries (QType-4, third upper subtable), and case-queries (QType-5, bottom subtable): best-performing values of precision, recall, and micro-averaged F-measure. (Bold values correspond to the best performance obtained by a competing method, for each query-set and evaluation criterion; LamBERTa performance values, formatted in italic, are also reported from Table 3.3 to ease the comparison with the competing methods)

	LamBERTa			*TextCNN*			*BiLSTM*			*TextRCNN*			*Seq2Seq-A*			*Transformer*		
i	P	R	F^{μ}	P	R	F^{μ}	P	R	F^{μ}	P	R	F^{μ}	P	R	F^{μ}	P	R	F^{μ}
Q_1	*0.975*	*0.962*	*0.961*	0.942	0.911	0.910	0.752	0.712	0.680	0.894	0.838	0.835	0.620	0.629	0.564	**0.959**	**0.950**	**0.948**
Q_2	*0.981*	*0.972*	*0.971*	0.959	0.940	0.936	0.735	0.730	0.681	0.894	0.852	0.845	0.630	0.670	0.585	**0.965**	**0.955**	**0.954**
Q_3	*0.994*	*0.990*	*0.990*	0.962	0.951	0.951	0.727	0.724	0.674	0.905	0.865	0.861	0.631	0.641	0.573	**0.970**	**0.965**	**0.965**
Q_4	*0.983*	*0.961*	*0.964*	0.971	0.933	0.938	0.789	0.772	0.736	0.919	0.875	0.872	0.727	0.703	0.665	**0.969**	**0.947**	**0.951**
Q_5	*0.944*	*0.910*	*0.909*	0.895	0.857	0.848	0.689	0.670	0.626	0.817	0.763	0.750	0.669	0.646	0.602	**0.936**	**0.903**	**0.903**
Q_6	*0.986*	*0.979*	*0.978*	0.949	0.938	0.935	0.745	0.723	0.677	0.903	0.866	0.858	0.757	0.718	0.685	**0.955**	**0.957**	**0.956**
Q_1	*0.866*	*0.841*	*0.828*	0.800	0.769	0.750	0.360	0.412	0.349	0.674	0.646	0.619	0.345	0.381	0.324	**0.801**	**0.774**	**0.754**
Q_2	*0.856*	*0.828*	*0.814*	0.760	0.726	0.706	0.409	0.442	0.382	0.646	0.643	0.601	0.353	0.384	0.327	**0.813**	**0.787**	**0.762**
Q_3	*0.886*	*0.861*	*0.851*	0.817	0.765	0.758	0.464	0.493	0.435	0.707	0.672	0.648	0.361	0.409	0.342	**0.852**	**0.804**	**0.795**
Q_4	*0.756*	*0.736*	*0.713*	0.729	0.643	0.669	0.310	0.330	0.292	0.643	0.607	0.582	0.278	0.297	0.258	**0.733**	**0.651**	**0.674**
Q_5	*0.759*	*0.718*	*0.710*	0.714	0.666	0.649	0.381	0.399	0.353	0.637	0.610	0.580	0.377	0.393	0.345	**0.734**	**0.715**	**0.697**
Q_6	*0.874*	*0.841*	*0.833*	0.788	0.737	0.725	0.451	0.451	0.405	0.662	0.625	0.603	0.378	0.438	0.364	**0.795**	**0.756**	**0.742**
Q_1	*0.197*	*0.190*	*0.170*	**0.158**	**0.152**	**0.124**	0.013	0.018	0.013	0.146	0.147	0.120	0.034	0.019	0.018	0.114	0.100	0.089
Q_2	*0.191*	*0.216*	*0.173*	**0.170**	0.155	**0.136**	0.015	0.037	0.009	0.166	**0.167**	0.133	0.011	0.034	0.012	0.121	0.093	0.092
Q_3	*0.223*	*0.241*	*0.204*	0.201	**0.208**	0.173	0.006	0.021	0.007	**0.205**	0.204	**0.176**	0.019	0.029	0.019	0.118	0.112	0.102
Q_4	*0.176*	*0.189*	*0.156*	0.159	0.149	0.126	0.006	0.007	0.005	**0.174**	**0.173**	**0.144**	0.006	0.016	0.004	0.073	0.071	0.061
Q_5	*0.132*	*0.092*	*0.090*	**0.109**	0.080	**0.090**	0.013	0.017	0.008	0.104	**0.092**	0.085	0.004	0.087	0.006	0.089	0.075	0.070
Q_6	*0.211*	*0.224*	*0.192*	0.183	0.166	0.155	0.014	0.024	0.014	**0.186**	**0.185**	**0.157**	0.017	0.065	0.020	0.120	0.103	0.090
Q_1	*0.233*	*0.228*	*0.210*	0.147	0.137	0.123	0.009	0.009	0.008	**0.155**	**0.144**	**0.123**	0.001	0.005	0.002	0.102	0.089	0.086
Q_2	*0.316*	*0.292*	*0.274*	**0.186**	0.163	0.132	0.009	0.023	0.009	0.172	**0.169**	**0.143**	0.023	0.012	0.016	0.118	0.092	0.104
Q_3	*0.323*	*0.284*	*0.273*	**0.221**	**0.267**	**0.219**	0.006	0.013	0.008	0.178	0.173	0.151	0.010	0.019	0.012	0.193	0.115	0.124
Q_4	*0.299*	*0.259*	*0.250*	**0.216**	**0.187**	**0.170**	0.007	0.009	0.004	0.115	0.112	0.095	0.021	0.011	0.011	0.106	0.088	0.097
Q_5	*0.401*	*0.354*	*0.342*	0.271	**0.251**	0.222	0.016	0.016	0.009	**0.288**	0.232	**0.226**	0.005	0.020	0.006	0.186	0.144	0.152
Q_6	*0.445*	*0.392*	*0.372*	**0.357**	**0.322**	**0.294**	0.045	0.050	0.033	0.301	0.303	0.259	0.022	0.029	0.016	0.201	0.130	0.137

Evaluation requirements. We designed this comparative evaluation to fulfill three main requirements: (i) significance of the selected competing methods, (ii) robustness of their evaluation, and (iii) relatively fair comparison among the competing methods and our LamBERTa models.

To meet the first requirement, we focused on deep neural network models that have been widely used for text classification and are often referred to as "predecessors" of deep pre-trained language models like BERT.

To ensure a robust evaluation, we carried out an extensive parameter-tuning phase of each competing methods, by varying all main parameters within recommended or reasonable range values, which include: dropout probability (within [0.1,0.4]), maximum sentence length (fixed to 60, or flexible), batch size (base-2 powers from 16 to 128), number of epochs (from 10 to 100). The results we will present here refer to the best performances achieved by each of the models.

We considered main aspects that are shared between our models and the competing ones to achieve a fair comparison. First, as each of the competing models, except the Transformer, needs word vector initialization, they were provided with Italian Wikipedia pre-trained Glove embeddings — recall that Italian Wikipedia is a major constituent of the pre-trained Italian BERT used for our LamBERTa models. Each model was then fine-tuned over the individual ICC books following the same data labeling schemes used for the LamBERTa models. Moreover, since we have to train a classification problem with as many classes as the number of articles for a given ICC book, all the models use an appropriate yet identical setting as in LamBERTa for the optimizer (i.e., Adam) and the loss function (i.e., cross entropy).

Experimental results. The goal of this evaluation stage was to demonstrate the superiority of LamBERTa against the selected competitors for case law retrieval as classification task, using different types of test queries. For each competitor, we used the individual books' `UniRR.T`$^{+}$ labeled data for training. We carried out several runs by varying the main parameters as previously discussed, and eventually, we selected the best-performing results for each competitor and query-set, which are shown in Table 3.10.

Looking at the table, there are a few remarks that stand out about the competitors. First, the *Transformer* model consistently excels over the other competitors in QType-1 and QType-2 query sets; this hints at a better effectiveness of the *Transformer* approach against queries that have some lexical affinity with the training text data. By contrast, when this does not hold as for comment (i.e., QType-4) and case law (i.e., QType-5) queries, *TextCNN* and *TextRCNN* tend to perform better than the others. In fact, for such types of queries, the lack of masked language modeling in *Transformer* seems not to be compensated by the attention mechanism

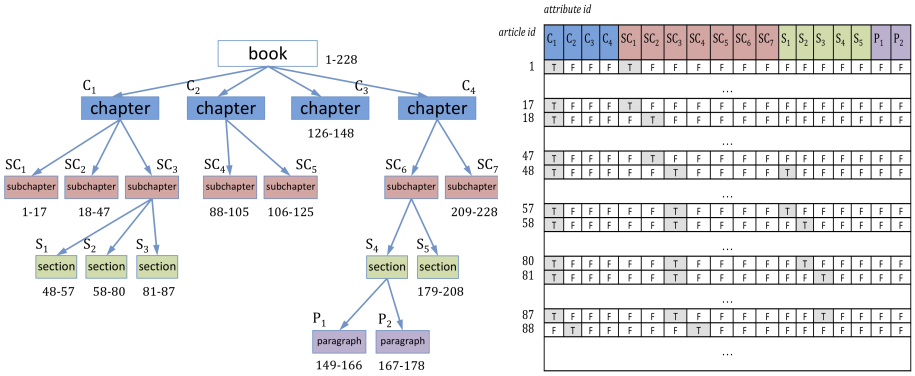


Fig. 3.3 Attribute-aware article prediction: Example book hierarchical organization (on the left) and excerpt of its corresponding attribute-aware article representation (on the right)

as it happens for the QType-1 and QType-2 queries. Moreover, the RNN models achieve lower scores than CNN-based or more advanced models, on all query sets. This indicates that the ability of *BiLSTM* and *Seq2Seq-A* of handling more distant token information (i.e., long range semantic dependency) and achieving complete abstraction at their bottom layers (without requiring multiple layer stacking like for CNNs) appear to vanish without a deep pre-training. In addition, CNN models are effective in extracting local and position-invariant features, and it has indeed been shown (e.g., [110]) that CNNs can successfully learn to classify a sentence (like QType-4 comment queries) or a paragraph (like QType-5 case law queries).

Our major finding is that, when comparing to the results obtained by LamBERTa models (cf. first column of Table 3.10), the best among the above competing methods turn out to be outperformed by the corresponding LamBERTa models, on all types of query. This confirms our initial expectation on the superiority of LamBERTa in learning classification models from few labeled examples per-class under a tough multi-class classification scenario, having a very large (i.e., in the order of hundreds) number of classes.

Attribute-aware law article prediction

As discussed in Section 1.3, the method in [40] was conceived for attribute-aware charge prediction, and was evaluated on collected criminal cases using two types of text data: the fact of each case and the charges extracted from the penalty results of each case. Moreover, a number of attributes were defined as to distinguish the confusing charge pairs previously selected from the confusion matrix obtained by a charge prediction model.

To apply the above method to the ICC law article prediction task, we define an *attribute-aware representation* of the ICC article data by exploiting the available ICC hierarchical

labeling of the articles of each book (cf. Section 2.4). Our objective is to leverage the ICC-classification based attributes as explicit knowledge about how to distinguish related groups of articles within the same book.

To this purpose, we adopt the following methodology. From the tree modeling the hierarchical organization into chapters, subchapters, sections and paragraphs of any given book, we treat each node as a boolean attribute and a complete path (from the chapter level to a leaf node) as an attribute-set assignment for each of the articles under that subtree path. We illustrate our defined procedure in Figure 3.3, where for the sake of simplicity, the example shows a hypothetical book containing 228 articles and organized into four chapters (i.e., $C_1, \ldots, C_4$), seven subchapters (i.e., $SC_1, \ldots, SC_7$), five sections (i.e., $S_1, \ldots, S_5$), and two paragraphs (i.e., P_1, P_2). Thus, eighteen attributes are defined in total, and each article in the book is associated with a binary vector (i.e., 1 means that an attribute is representative for the article, 0 otherwise); note also that the bunch of articles within the same hierarchical level subdivision share the attribute representation.

The ICC-heading attributes resulting from the application of the above procedure on each of the ICC books are reported in Appendix, Tables 3.16–3.21.

Experimental settings and results. We resorted to the software implementation of the method in [40] provided by the authors[10] and made minimal adaptations to the code so to enable the method to work with our ICC data. We hereinafter refer to this method as *attribute-aware few-shot article prediction* model, *A-FewShotAP*.

Analogously to the previous stage of comparative evaluation, we conducted an extensive parameter-tuning phase of *A-FewShotAP*, by considering both default values and variations for the main parameters, such as number of layers, learning rate, number of epochs, batch size. Moreover, like for the other competitors, we used Italian Wikipedia pre-trained embeddings.

Table 3.11 summarizes the best-performing results obtained by *A-FewShotAP* over different types of query sets, namely paraphrased-sentence-queries, comment-sentence-queries, and case-queries.

As it can be observed, *A-FewShotAP* performance scores reveal to be never comparable to those produced by our LamBERTa models. Indeed, despite no information on article attributes is exploited in LamBERTa models, the latter achieve a large effectiveness gain over *A-FewShotAP*, regardless of the type of query.

Nonetheless, it is also worth noticing the beneficial effect of the attribute-aware article prediction when comparing *A-FewShotAP* with other RNN-based competing methods, particularly *BiLSTM* and *Seq2Seq-A*. Notably, although *A-FewShotAP* builds on a conventional LSTM — unlike *BiLSTM* and *Seq2Seq-A*, which utilize a bidirectional LSTM — it can

[10] https://github.com/thunlp/attribute_charge

Table 3.11 Attribute-aware article prediction: The *A-FewShotAP* model [40] trained with each ICC book annotated with the UniRR.T$^+$ labeling scheme, and tested over all sets of paraphrased-sentence-queries (QType-2, upper subtable), comment-sentence-queries (QType-4, second upper subtable), and case-queries (QType-5, bottom subtable). The table shows best-performing values of precision, recall, and macro-averaged F-measure. (LamBERTa performance values, formatted in italic, are also reported from Table 3.3 to ease the comparison with the competing method)

	LamBERTa			*A-FewShotAP* [40]		
i	P	R	F^M	P	R	F^M
Q_1	*0.866*	*0.841*	*0.853*	0.355	0.410	0.381
Q_2	*0.856*	*0.828*	*0.841*	0.328	0.378	0.351
Q_3	*0.886*	*0.861*	*0.873*	0.450	0.514	0.480
Q_4	*0.756*	*0.736*	*0.746*	0.319	0.384	0.348
Q_5	*0.759*	*0.718*	*0.738*	0.382	0.444	0.411
Q_6	*0.874*	*0.841*	*0.857*	0.412	0.474	0.441
Q_1	*0.197*	*0.190*	*0.194*	0.018	0.031	0.023
Q_2	*0.191*	*0.216*	*0.203*	0.023	0.043	0.030
Q_3	*0.223*	*0.241*	*0.231*	0.027	0.038	0.032
Q_4	*0.176*	*0.189*	*0.182*	0.029	0.047	0.036
Q_5	*0.132*	*0.092*	*0.108*	0.035	0.042	0.038
Q_6	*0.211*	*0.224*	*0.218*	0.031	0.037	0.034
Q_1	*0.233*	*0.228*	*0.230*	0.006	0.013	0.009
Q_2	*0.316*	*0.292*	*0.303*	0.016	0.021	0.018
Q_3	*0.323*	*0.284*	*0.302*	0.016	0.018	0.017
Q_4	*0.299*	*0.259*	*0.278*	0.007	0.010	0.008
Q_5	*0.401*	*0.354*	*0.376*	0.018	0.027	0.022
Q_6	*0.445*	*0.392*	*0.417*	0.025	0.037	0.030

achieve comparable or even higher performance than the other two methods (cf. Table 3.10): in fact, on QType-4 queries, *A-FewShotAP* behaves better than *BiLSTM* and (especially in terms of precision) than *Seq2Seq-A*, while it performs better than *Seq2Seq-A* and is generally as good as *BiLSTM* on QType-2 and QType-5 queries.

3.4 Rebuilding LamBERTa based on ITALIAN-LEGAL-BERT

In this section we aim to answer the following research questions:

- **RQ1**: How does the behavior of LamBERTa models change when fine-tuning a legal Italian BERT rather than a general-domain Italian BERT?

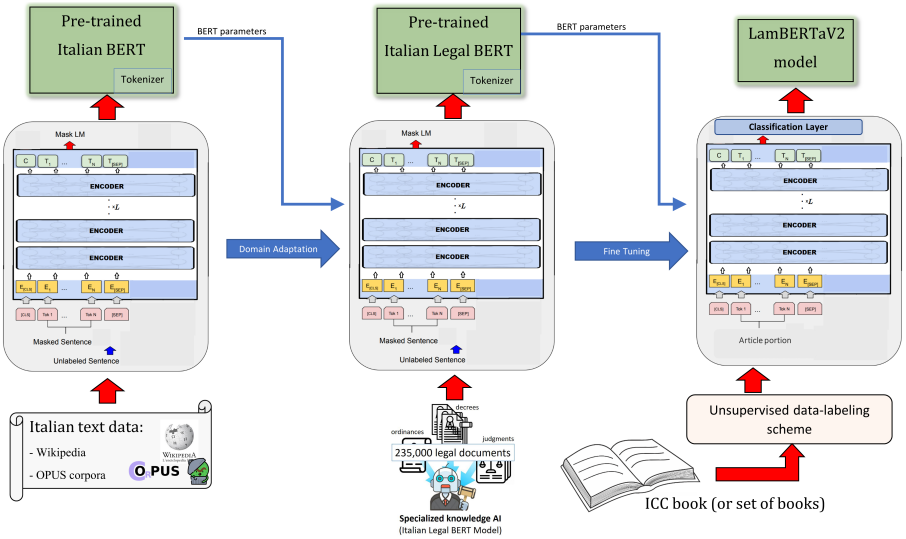


Fig. 3.4 An illustration of the conceptual architecture of LamBERTa-V2 designed for the ICC law article retrieval task.

- **RQ2**: What is the impact of injecting out-of-vocabulary legal terms into LamBERTa models during the fine-tuning stage? Does it depend on how such terms' representation is initialized?
- **RQ3**: What aspects arise from the explanation of the different LamBERTa models through the interpretation of their predictions?
- **RQ4**: Overall, is the combined effect of domain-adaptation and task-adaptation of a pre-trained Italian BERT model helpful to improve performance on the task of article retrieval from the Italian Civil Code?

To answer our first research question (**RQ1**), we develop a new version of LamBERTa by replacing the general-domain pre-trained Italian model (i.e., ITALIAN-XXL-UNCASED) with a legal-specific pre-trained Italian model (i.e., ITALIAN-LEGAL-BERT); recall that the latter model is the result of a further pre-training of the former, although on a cased version. In this section we refer to this version of LamBERTa as LamBERTa-V2, to distinguish from the original denoted as LamBERTa-V1. Furthermore, we provide background concepts on the ITALIAN-LEGAL-BERT [56]. Figure 3.4 shows the conceptual architecture of our proposed LamBERTa-V2.

3.4.1 ITALIAN-LEGAL-BERT

Table 3.12 LamBERTa-V1 vs. LamBERTa-V2 for all books of the ICC on book-sentence-queries (QT1), paraphrased-sentence-queries (QT2), comment-queries (QT3), and case-queries (QT5): Recall, Precision, micro- and macro-averaged F-measures, Recall@k, and MRR. (*Bold values correspond to the best model for each book, evaluation criterion, and query set*)

		R		P		F^{μ}		F^{M}		R@3		R@10		MRR	
Query type	Book	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
QT1	B_1	0.962	**0.975**	0.975	**0.983**	0.961	**0.973**	0.969	**0.979**	0.989	**0.997**	0.999	**1.000**	0.976	**0.983**
	B_2	0.972	0.972	0.981	**0.982**	0.971	**0.972**	0.977	0.977	0.994	0.994	**0.999**	0.997	0.983	0.983
	B_3	**0.990**	0.935	**0.994**	0.961	**0.990**	0.936	**0.992**	0.948	**1.000**	0.972	**1.000**	0.985	**0.995**	0.957
	B_4	**0.961**	0.665	**0.983**	0.768	**0.964**	0.672	**0.972**	0.712	**0.984**	0.750	**0.990**	0.831	**0.975**	0.721
	B_5	**0.910**	0.562	**0.944**	0.697	**0.909**	0.581	**0.927**	0.622	**0.984**	0.687	**0.998**	0.769	**0.947**	0.639
	B_6	0.979	0.979	0.986	0.986	0.978	0.978	0.982	**0.983**	1.000	1.000	1.000	1.000	0.989	0.989
QT2	B_1	**0.841**	0.815	**0.866**	0.837	**0.828**	0.800	**0.853**	0.826	0.905	**0.909**	**0.949**	0.939	**0.881**	0.865
	B_2	**0.828**	0.799	**0.856**	0.842	**0.814**	0.788	**0.841**	0.820	**0.892**	0.881	**0.941**	0.919	**0.871**	0.845
	B_3	**0.861**	0.740	**0.886**	0.773	**0.851**	0.725	**0.873**	0.756	**0.922**	0.831	**0.942**	0.870	**0.896**	0.794
	B_4	**0.736**	0.463	**0.756**	0.477	**0.713**	0.436	**0.746**	0.470	**0.806**	0.528	**0.861**	0.598	**0.779**	0.512
	B_5	**0.718**	0.401	**0.759**	0.451	**0.710**	0.393	**0.738**	0.425	**0.843**	0.508	**0.908**	0.582	**0.790**	0.472
	B_6	0.841	**0.852**	0.874	**0.886**	0.833	**0.843**	0.857	**0.869**	0.914	**0.918**	0.941	**0.947**	0.882	**0.889**
QT3	B_1	**0.349**	0.297	**0.248**	0.192	**0.274**	0.222	**0.290**	0.233	**0.494**	0.462	**0.675**	0.636	**0.455**	0.411
	B_2	0.313	**0.320**	0.213	**0.226**	0.239	**0.251**	0.253	**0.265**	0.494	**0.522**	0.655	**0.661**	**0.445**	0.440
	B_3	**0.396**	0.260	**0.298**	0.195	**0.325**	0.212	**0.340**	0.223	**0.577**	0.376	**0.704**	0.518	**0.507**	0.346
	B_4	**0.336**	0.078	**0.239**	0.059	**0.264**	0.064	**0.279**	0.067	**0.487**	0.115	**0.622**	0.160	**0.438**	0.107
	B_5	**0.171**	0.019	**0.128**	0.018	**0.137**	0.018	**0.147**	0.019	**0.364**	0.037	**0.562**	0.073	**0.302**	0.041
	B_6	0.387	**0.442**	0.291	**0.334**	0.317	**0.364**	0.332	**0.381**	0.588	**0.619**	0.745	**0.765**	0.515	**0.549**
QT5	B_1	**0.228**	0.182	**0.233**	0.194	**0.210**	0.164	**0.230**	0.187	**0.494**	0.360	**0.813**	0.526	**0.430**	0.328
	B_2	**0.292**	0.246	**0.316**	0.269	**0.274**	0.217	**0.303**	0.257	**0.501**	0.447	**0.726**	0.651	**0.465**	0.396
	B_3	**0.284**	0.197	**0.323**	0.220	**0.273**	0.177	**0.302**	0.208	**0.566**	0.289	**0.856**	0.436	**0.489**	0.268
	B_4	**0.259**	0.050	**0.299**	0.079	**0.250**	0.055	**0.278**	0.061	**0.528**	0.108	**0.803**	0.148	**0.458**	0.092
	B_5	**0.354**	0.019	**0.401**	0.046	**0.342**	0.023	**0.376**	0.027	**0.641**	0.040	**0.884**	0.062	**0.564**	0.040
	B_6	**0.392**	0.378	**0.445**	0.376	**0.372**	0.334	**0.417**	0.377	**0.665**	0.577	**0.885**	0.720	**0.590**	0.514

ITALIAN-LEGAL-BERT [56] follows the typical BERT architecture, with a language modeling head on top, AdamW Optimizer, initial learning rate 5e-5. ITALIAN-LEGAL-BERT was built upon ITALIAN-XXL-CASED by further pre-training the latter for additional 4 epochs on a corpus extracted from the National Jurisprudential Archive (pst.giustizia.it), a repository containing millions of legal documents, such as decrees, orders, and civil judgments, from Italian courts and courts of appeal. The corpus used to train ITALIAN-LEGAL-BERT is 3.7 GB of text containing above 21M sentences and 498M words. The trained ITALIAN-LEGAL-BERT was evaluated on named entity recognition, sentence classification, and sentence similarity tasks, using 20K civil cases from the National Jurisprudential Archive and above 21K criminal cases from italgiureweb (italgiure.giustizia.it).

Table 3.12 summarizes results of the comparison between the two versions based on their evaluation through all query sets.[11]

At a first glance, it can be noticed that although there is no absolute winner, LamBERTa-V1 generally achieves better performance than LamBERTa-V2. For all query types, LamBERTa-V2 appears to lose more when evaluated on queries pertaining the largest books (i.e., Book-4 and Book-5). Moreover, regardless of the book, the gap of LamBERTa-V2 is particularly evident for the most difficult query sets, i.e., QT3 and QT5, which contain queries that are the most distant, both lexically and semantically, from the language used in the training instances. On average over all books, LamBERTa-V2 has indeed a percentage decrease of above 40% on case queries (QT5) and above 27% on comment queries (QT3); remarkably, while this holds for all criteria, the negative peaks are reached for the top-3 and top-10 predictions: -46.4% $R@3$ and -48.8% $R@10$, on QT5, and about 29% $R@3$ and $R@10$, on QT3. In light of the above results, it stands out that using a legal pre-trained model does not bring advantage over a general-domain pre-trained model to fine-tune on the downstream task of ICC article prediction, and actually the legal pre-trained model can often achieve worse performance.

3.4.2 Investigating on the domain-specific token injection

Effect of token injection removal. A major goal of this work is to delve into the effect of the domain-specific token injection into LamBERTa models. To answer our **RQ2**, we first analyze the changes in the behavior of LamBERTa when no out-of-vocabulary tokens are added. We shall use suffix NoDST to distinguish this setting from the original one using token injection.

Results obtained by LamBERTa-V1-NoDST and LamBERTa-V2-NoDST are shown in Table 3.13. First, we notice that the performance difference of LamBERTa-V2-NoDST w.r.t. LamBERTa-V1-NoDST is reduced, though still remaining negative, with the exception of QT5, where LamBERTa-V2-NoDST achieves average percentage increase of about 3% P up to 9% R. More interesting is to compare the obtained results against those in Table 3.12. The new setting leads to an improvement of both versions of LamBERTa in most cases, where the ITALIAN-LEGAL-BERT based version takes major benefits. More precisely, the two versions of LamBERTa improves slightly on QT1 and QT2, and more significantly on QT3. By contrast, on QT5, while LamBERTa-V2-NoDST achieves average percentage increase vs. LamBERTa-V2 (from 65% to about 90%), taking light advantage on other models in terms of

[11] All results shown in Tables 3.12–3.14 correspond to the use of same seed setting (for handling computation randomness) and hardware configuration for all LamBERTa models.

Table 3.13 LamBERTa-V1-NoDST vs. LamBERTa-V2-NoDST for all books of the ICC on book-sentence-queries (QT1), paraphrased-sentence-queries (QT2), comment-queries (QT3), and case-queries (QT5): Recall, Precision, micro- and macro-averaged F-measures, Recall@k, and MRR. (*Bold values correspond to the best model for each book, evaluation criterion, and query set*)

		R		P		F^{μ}		F^{M}		$R@3$		$R@10$		MRR	
Query type	Book	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
QT1	B_1	0.975	0.975	0.983	0.983	0.973	0.973	0.979	0.979	**0.996**	0.995	1.000	1.000	0.985	0.985
	B_2	0.974	**0.975**	0.983	**0.985**	0.973	**0.975**	0.978	**0.980**	0.993	**0.994**	0.996	**0.999**	0.984	**0.985**
	B_3	0.988	0.988	**0.991**	0.990	0.987	0.987	0.989	0.989	1.000	1.000	1.000	1.000	0.995	0.995
	B_4	**0.968**	0.967	**0.989**	0.988	**0.971**	0.970	**0.978**	0.977	0.985	0.985	0.990	0.990	0.977	0.977
	B_5	**0.924**	0.922	**0.957**	0.953	**0.924**	0.920	**0.941**	0.937	**0.986**	0.985	**0.999**	0.998	**0.956**	0.954
	B_6	0.979	**0.980**	0.986	**0.987**	0.978	**0.979**	0.983	**0.984**	1.000	1.000	1.000	1.000	0.989	**0.990**
QT2	B_1	**0.863**	0.851	**0.875**	0.874	**0.849**	0.839	**0.869**	0.862	**0.929**	0.920	**0.962**	0.957	**0.901**	0.890
	B_2	**0.867**	0.845	**0.882**	0.859	**0.854**	0.830	**0.874**	0.852	**0.920**	0.916	**0.959**	0.945	**0.900**	0.885
	B_3	**0.890**	0.867	**0.898**	0.875	**0.876**	0.852	**0.894**	0.871	**0.939**	0.917	**0.963**	0.943	**0.920**	0.898
	B_4	**0.761**	0.753	**0.774**	0.769	**0.738**	0.730	**0.767**	0.760	**0.834**	0.833	0.882	**0.889**	**0.804**	0.801
	B_5	**0.816**	0.778	**0.838**	0.791	**0.801**	0.755	**0.827**	0.785	**0.918**	0.886	**0.952**	0.934	**0.870**	0.839
	B_6	**0.867**	0.864	0.884	**0.896**	0.856	0.856	0.875	**0.880**	0.931	0.931	**0.962**	0.960	**0.904**	0.900
QT3	B_1	**0.437**	0.388	**0.318**	0.278	**0.351**	0.308	**0.368**	0.324	**0.612**	0.575	**0.740**	0.731	**0.547**	0.510
	B_2	**0.491**	0.478	**0.368**	0.360	**0.401**	0.394	**0.421**	0.411	**0.652**	0.643	0.792	**0.807**	**0.596**	0.589
	B_3	**0.551**	0.477	**0.444**	0.371	**0.475**	0.403	**0.492**	0.418	**0.709**	0.649	**0.805**	0.787	**0.648**	0.591
	B_4	0.447	**0.449**	**0.349**	0.343	**0.377**	0.372	**0.392**	0.389	0.602	**0.642**	0.720	**0.770**	0.545	**0.566**
	B_5	**0.401**	0.322	**0.301**	0.241	**0.327**	0.260	**0.344**	0.276	**0.567**	0.537	**0.712**	0.697	**0.513**	0.460
	B_6	0.497	**0.553**	0.396	**0.443**	0.425	**0.474**	0.441	**0.492**	0.673	**0.707**	0.803	**0.833**	0.600	**0.653**
QT5	B_1	**0.190**	0.175	0.201	**0.210**	**0.174**	0.171	**0.195**	0.191	**0.367**	0.353	**0.581**	0.529	**0.347**	0.313
	B_2	0.351	**0.357**	0.362	**0.375**	0.318	**0.327**	0.356	**0.365**	0.539	**0.562**	0.697	**0.720**	0.498	**0.502**
	B_3	0.325	**0.329**	**0.354**	0.323	**0.303**	0.291	**0.339**	0.326	0.518	**0.526**	0.675	**0.678**	0.443	**0.455**
	B_4	0.259	**0.301**	**0.297**	0.295	0.245	**0.271**	0.276	**0.298**	0.416	**0.510**	0.577	**0.670**	0.372	**0.446**
	B_5	**0.406**	0.371	**0.421**	0.372	**0.379**	0.343	**0.413**	0.371	**0.650**	0.647	0.768	**0.802**	0.588	0.558
	B_6	0.307	**0.470**	0.352	**0.470**	0.297	**0.429**	0.328	**0.470**	0.544	**0.659**	0.698	**0.816**	0.474	**0.600**

R, P, F criteria, LamBERTa-V1-NoDST tends to be worse than the original LamBERTa-V1 which remains the absolute winner according to the $R@k$ and MRR criteria.

Effect of embedding initialization for token injection. The above results prompted us to further investigate on the effect of injecting out-of-vocabulary legal terms into LamBERTa models, by focusing now on the *initialization* of the added tokens. In fact, it should be noted that in the original setting of LamBERTa, the selected domain-specific tokens are added to the Italian pre-trained tokenizer using a random initialization. Therefore, to provide a more exhaustive answer to our **RQ2**, we define an enhanced setting for the domain-specific tokens to be added. Our goal is to compute initial embeddings for the new tokens that are not random but incorporate proper knowledge of the legal language. One approach we tried is to initialize each word w to be added by getting the [CLS] output embedding computed when prompting the Italian pre-trained model, or alternatively ITALIAN-LEGAL-BERT, with just

w. Similarly, we tried by averaging the output embeddings of the tokens corresponding to the subwords of *w* detected by the BERT tokenizer. Unfortunately, in both cases, exploiting the output embeddings shows to be inappropriate, which might be due to the fact that these contextualized representations incorporate also the segment and position embeddings. Then, we shifted our attention to vectors extracted from the token embeddings matrix (i.e., the first level of BERT input representation). Given a word *w* to be added, we get the token embedding of each of its subwords (excluding [CLS] and [SEP] embeddings) and tried different pooling strategies. The one leading to the best results is initializing the *w*'s embedding with the initial embedding of its root subword. This setting is hereinafter referred to as ReDST.

Results for this new setting are reported in Table 3.14. A first remark that stands out is the benefit brought by this new setting of initialization of the injected tokens w.r.t. a random initialization, for both versions of LamBERTa. This holds always, with the exception of QT4 according to $R@k$ and *MRR* criteria, whereby LamBERTa-V1 is the absolute winner over all models. Besides, LamBERTa-V1-ReDST and LamBERTa-V2-ReDST actually perform comparably or better than LamBERTa-V1-NoDST and LamBERTa-V2-NoDST on QT1, and clearly better than LamBERTa-V1-NoDST and LamBERTa-V2-NoDST on QT5 according to $R@k$ and *MRR*.

3.5 Explainability

Like for any machine and deep learning models, *explainability* of PLMs is central to understand their solutions provided for a given NLP task. This becomes even more crucial when artificial intelligence meets a challenging field like law (e.g., [14, 38]). We start our understanding of what is going on inside the "black box" of LamBERTa models. One important aspect is the *explainability* of LamBERTa models, which we investigate first focusing on how they form complex relationships between the textual tokens, and their distinctive attention patterns. Then, we take a different perspective, which is more suited for providing interpretation of the models' prediction. To this purpose, we use LIME - Local Interpretable Model-Agnostic Explanations [86].

3.5.1 Explainability of LamBERTa models based on Attention Patterns

Like any BERT-based architecture, our LamBERTa leverages the Transformer paradigm, which allows for processing all elements simultaneously by forming direct connections between individual elements through a mechanism known as *attention*. Attention is a way for a model to assign weight to input features (i.e., parts of the texts) based on their high

Table 3.14 LamBERTa-V1-ReDST vs. LamBERTa-V2-ReDST for all books of the ICC on book-sentence-queries (QT1), paraphrased-sentence-queries (QT2), comment-queries (QT3), and case-queries (QT5): Recall, Precision, micro- and macro-averaged F-measures, Recall@k, and MRR. (*Bold values correspond to the best model for each book, evaluation criterion, and query set*)

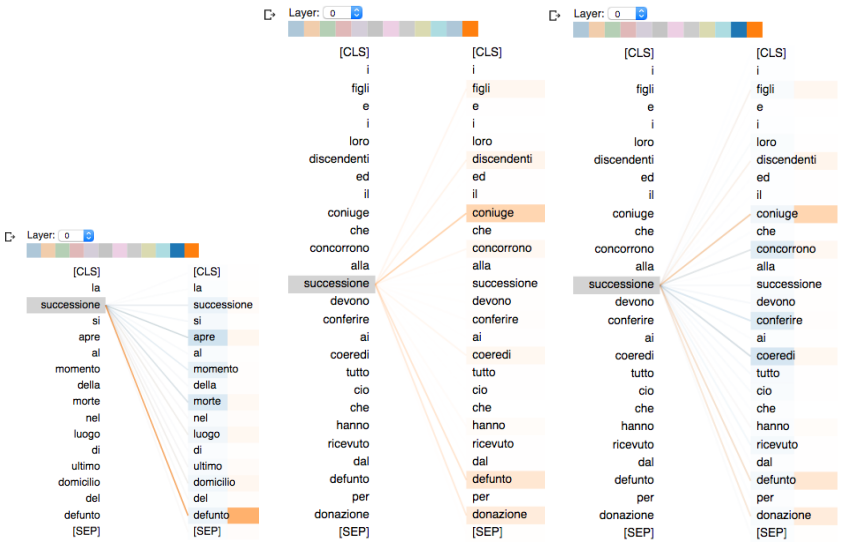
		R		P		F^{μ}		F^{M}		R@3		R@10		MRR	
Query type	Book	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
QT1	B_1	**0.973**	0.972	0.981	**0.982**	**0.972**	0.971	0.977	0.977	0.996	0.996	1.000	1.000	0.984	0.984
	B_2	0.974	0.974	0.983	0.983	0.973	0.973	0.978	0.978	0.994	0.994	0.997	0.997	0.984	0.984
	B_3	0.988	0.988	0.991	0.991	0.987	0.987	0.989	0.989	1.000	1.000	1.000	1.000	0.995	0.995
	B_4	0.968	0.968	0.989	0.989	0.971	0.971	0.978	0.978	0.985	0.985	0.990	0.990	0.977	0.977
	B_5	**0.924**	0.920	**0.957**	0.952	**0.924**	0.919	**0.941**	0.936	**0.986**	0.985	0.998	**0.999**	**0.955**	0.953
	B_6	0.979	0.979	0.986	0.986	0.978	0.978	0.983	0.983	1.000	1.000	1.000	1.000	0.989	0.989
QT2	B_1	**0.846**	0.843	0.861	**0.870**	0.832	**0.833**	0.853	**0.856**	**0.928**	0.918	**0.957**	0.951	**0.890**	0.884
	B_2	**0.843**	0.829	**0.859**	0.850	**0.827**	0.814	**0.851**	0.839	**0.913**	0.904	**0.951**	0.943	**0.885**	0.870
	B_3	**0.883**	0.854	**0.889**	0.871	**0.868**	0.841	**0.886**	0.862	**0.924**	0.909	**0.958**	0.939	**0.911**	0.887
	B_4	**0.763**	0.724	**0.778**	0.729	**0.739**	0.696	**0.770**	0.727	**0.840**	0.804	**0.886**	0.864	**0.808**	0.772
	B_5	**0.792**	0.763	**0.813**	0.783	**0.773**	0.742	**0.802**	0.773	**0.898**	0.872	**0.944**	0.921	**0.850**	0.825
	B_6	0.844	**0.849**	0.875	**0.877**	0.834	**0.838**	0.860	**0.863**	0.921	0.921	0.951	**0.957**	0.887	**0.888**
QT3	B_1	**0.408**	0.364	**0.308**	0.266	**0.334**	0.292	**0.351**	0.308	**0.578**	0.498	**0.709**	0.670	**0.518**	0.463
	B_2	**0.450**	0.335	**0.344**	0.230	**0.374**	0.259	**0.390**	0.273	**0.637**	0.540	**0.786**	0.702	**0.564**	0.466
	B_3	**0.477**	0.372	**0.379**	0.289	**0.407**	0.311	**0.423**	0.325	**0.642**	0.525	**0.745**	0.695	**0.581**	0.477
	B_4	**0.439**	0.353	**0.341**	0.241	**0.370**	0.272	**0.384**	0.287	**0.594**	0.514	**0.695**	0.663	**0.536**	0.460
	B_5	**0.415**	0.316	**0.312**	0.238	**0.339**	0.258	**0.356**	0.272	**0.575**	0.484	**0.724**	0.636	**0.521**	0.432
	B_6	**0.507**	0.433	**0.406**	0.344	**0.434**	0.365	**0.451**	0.383	**0.656**	0.653	0.813	**0.823**	**0.612**	0.567
QT5	B_1	**0.181**	0.165	**0.212**	0.181	**0.177**	0.152	**0.195**	0.172	**0.351**	0.338	**0.532**	0.523	**0.335**	0.309
	B_2	0.296	**0.299**	**0.336**	0.329	**0.276**	0.274	**0.315**	0.313	0.530	**0.539**	0.677	**0.715**	**0.471**	0.464
	B_3	**0.298**	0.258	**0.298**	0.223	**0.262**	0.218	**0.298**	0.239	**0.496**	0.441	**0.653**	0.629	**0.443**	0.395
	B_4	**0.290**	0.267	**0.291**	0.280	**0.259**	0.244	**0.291**	0.273	0.455	**0.461**	**0.613**	0.597	**0.406**	0.404
	B_5	**0.323**	0.295	**0.337**	0.314	**0.298**	0.271	**0.330**	0.304	**0.620**	0.575	**0.774**	0.740	**0.526**	0.496
	B_6	**0.430**	0.351	**0.471**	0.369	**0.417**	0.329	**0.450**	0.359	**0.658**	0.551	**0.782**	0.747	**0.587**	0.499

informativeness or importance to some task. In particular, attention enables the model to understand how the words relate to each other in the context of the sentence, by forming composite representations that the model can reason about.

BERT's attention patterns can assume several forms, such as delimiter-focused, bag-of-words, and next-word patterns. Through the lens of the `bertviz` visualization tool,[12] we show how LamBERTa forms its distinctive attention patterns. For this purpose, here we present a selection of examples built upon sentences from Book-2 of the ICC (i.e., relevant to key concepts in inheritance law), which are next reported both in Italian and English-translated versions.

The attention-head view in `bertviz` visualizes attention patterns as lines connecting the word being updated (left) with the word(s) being attended to (right), for any given

[12] https://github.com/jessevig/bertviz.



(a) (b) (c)

Fig. 3.5 Attention patterns for "succession": (a) two-head pattern from Art. 456 and comparison with (b) single-head and (c) two-head patterns from Art. 737.

input sequence, where color intensity reflects the attention weight. Figure 3.5(a)-(c) shows noteworthy examples focused on the word "succession", from the following sentences:

> from Art. 456: *"la successione si apre al momento della morte nel luogo dell'ultimo domicilio del defunto"* (i.e., "the succession opens at the moment of death in the place of the last domicile of the deceased person")
>
> from Art. 737: *"i figli e i loro discendenti ed il coniuge che concorrono alla successione devono conferire ai coeredi tutto ciò che hanno ricevuto dal defunto per donazione"* (i.e., the children and their descendants and the spouse who contribute to the succession must give to the co-heirs everything they have received from the deceased person as a donation)

In Fig. 3.5(a), we observe how the source is connected to a meaningful, non-contiguous set of words, particularly, "apre" ("opens"), "morte" ("death"), and "defunto" ("deceased person"). In addition, in Fig. 3.5(b), we observe how "successione" is related to "coniuge" ("spouse"), "donazione" ("donation"), which is further enriched in the two-head attention patterns with "coeredi" ("co-heirs"), "conferire" ("give"), and "concorrono" ("contribute"), shown in Fig. 3.5(c); moreover, "successione" is still connected to "defunto". Remarkably, these patterns highlight the model's ability not only to mine semantically meaningful patterns that are more complex than next-word or delimiter-focused patterns, but also to build patterns that consistently hold across various sentences sharing words. Note that, as shown in our examples, these sentences can belong to different contexts (i.e., different articles), and can significantly vary in length. The latter point is particularly evident, for instance, in the following example sentences:

from Art. 457: *"l'eredità si devolve per legge o per testamento. Non si fa luogo alla successione legittima se non quando manca, in tutto o in parte, quella testamentaria"* (i.e., "the inheritance is devolved by law or by will. There is no place for legitimate succession except when the testamentary succession is missing, in whole or in part")

from Art. 683: *"la revocazione fatta con un testamento posteriore conserva la sua efficacia anche quando questo rimane senza effetto perché l'erede istituito o il legatario è premorto al testatore, o è incapace o indegno, ovvero ha rinunziato all'eredità o al legato"* (i.e., " the revocation made with a later will retains its effectiveness even when this remains without effect because the established heir or legatee is premortal to the testator, or is incapable or unworthy, or has renounced the inheritance or the legacy")

Focusing now on "eredità" ("inheritance"), in Art. 457 we found attention patterns with "testamento" ("testament"), "legittima" ("legitimate"), "testamentaria" ("testamentary succession"), whereas in the long sentence from Art. 683, we again found a link to "testamento" ("testament") as well as with the related concept "testatore" ("testator"), in addition to connections with "premorto" ("premortal"), "indegno" ("unworthy"), and "rinunziato" ("renounced").

We point out that our investigation of LamBERTa models' attention patterns was found to be persistent over different input sequences, besides the above discussed examples. This has unveiled the ability of LamBERTa models to form different types of patterns, which include complex bag-of-words and next-word patterns.

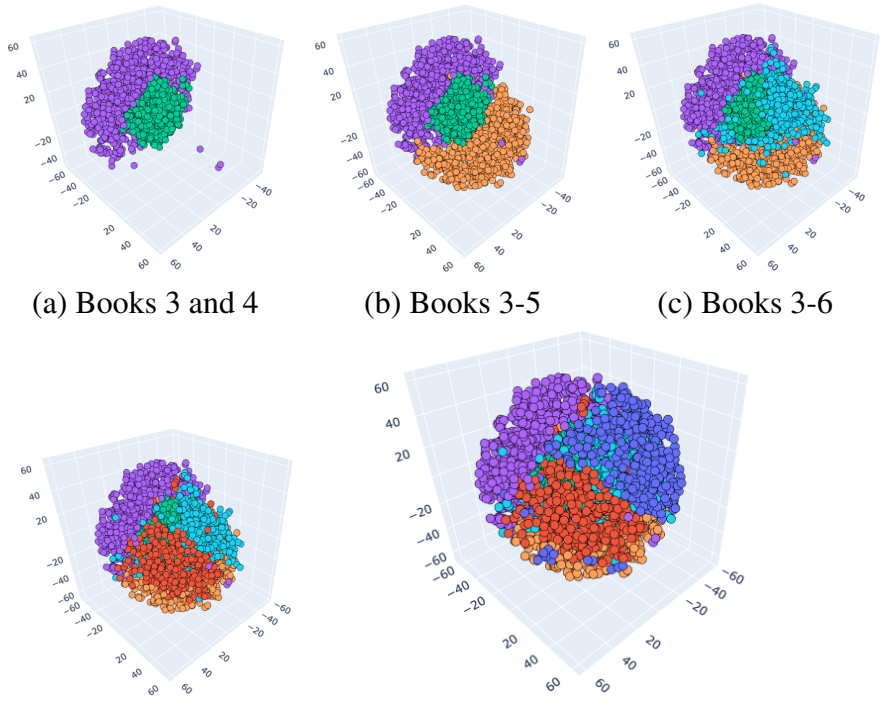
3.5.2 Visualization of ICC LamBERTa Embeddings

Our qualitative evaluation of LamBERTa models is also concerned with the core outcome of the models, that is, the learned representation embeddings for the inputted text.

As previously mentioned in Section 3.1.6, for any given text in input, a real-valued vector C of length 768 is produced downstream of the 12 layers of a BERT model, in correspondence with the input embedding $E_{[CLS]}$ relating to the special token CLS. This output embedding C can be seen as an encoded representation of the input text supplied to the model.

In the following, we present a visual exploratory analysis of LamBERTa embeddings based on a powerful, widely-recognized method for the visualization of high-dimensional data, namely *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [99]. t-SNE is a non-linear technique for dimensionality reduction that is highly effective at providing an intuition of how the data is arranged in a high-dimensional space. The t-SNE algorithm calculates a similarity measure between pairs of instances in the high-dimensional space and in the low-dimensional space, usually 2D or 3D. Then, it converts the pairwise similarities to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

It is well-known that t-SNE outputs provide better and more interpretable results than Principal Component Analysis (PCA) and other linear dimensionality reduction models,



(a) Books 3 and 4 (b) Books 3-5 (c) Books 3-6

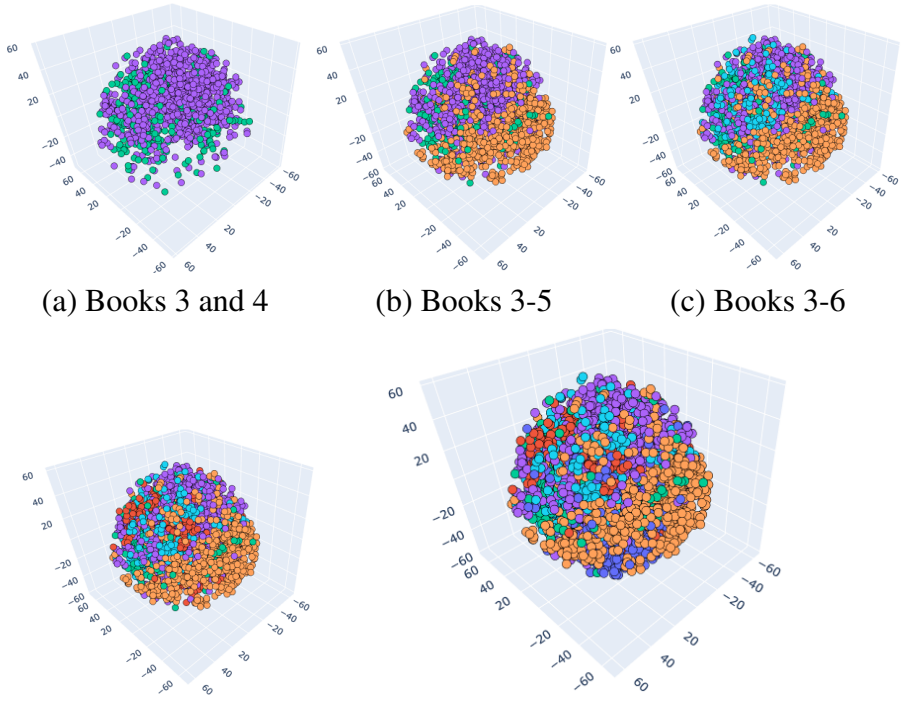
(d) Books 2-6 (e) All ICC books

Fig. 3.6 Visualization of the ICC article embeddings produced by LamBERTa local models, transformed onto a 3D t-SNE space. Color codes correspond to ICC books: blue for Book-1, red for Book-2, green for Book-3, purple for Book-4, orange for Book-5, cyan for Book-6

which are not effective at interpreting complex polynomial relationships between features. In particular, by seeking to maximize variance and preserving large pairwise distances, the classic PCA focuses on placing dissimilar data points far apart in a lower dimension representation; however, in order to represent high-dimensional data on low-dimensional, non-linear manifold, it is important that similar data points must be represented close together. This is ensured by t-SNE, which preserves small pairwise distances or local similarities (i.e., nearest-neighbors), so that similar points on the manifold are mapped to similar points in the low-dimensional representation.

Figure 3.6 displays 3D t-SNE representations of the article embeddings generated by LamBERTa local models, for each book of the ICC. (Each point represents the t-SNE transformation of the embedding of an article onto a 3D space, while colors are used to distinguish the various books.) In figure, we show progressive combinations of the results from different books,[13] while the last chart corresponds to the whole ICC. This choice of presentation is motivated to ease the readability of each book's article embeddings. It can be noted that the LamBERTa local models are able to generate embeddings so that t-SNE can

[13]The order of combination of the books is just due to the sake of presentation.



(a) Books 3 and 4 (b) Books 3-5 (c) Books 3-6

(d) Books 2-6 (e) All ICC books

Fig. 3.7 Visualization of the ICC article embeddings produced by LamBERTa global model, transformed onto a 3D t-SNE space. Color coding is the same as in Fig. 3.6

effectively put nearest-neighbor cases together, with a certain tendency of distributing the points from different books in differest subspaces.

Let us now compare the above results with those from Figure 3.7, which shows the 3D t-SNE representations of the article embeddings generated by LamBERTa global model.[14] From the comparison, we observe a less compact and localized representation in the global model embeddings w.r.t. the local model ones. This is interesting yet expected, since global models are designed to go beyond the boundaries of a book, whereas local models focus on the interrelations between articles on the same book.

It should be noted that while the above remark would hint at preferring the use of local models against global models, at least in terms of interpretability based on visual exploratory analysis, we shall deep our understanding on the effectiveness of the two learning approaches in the next section, which is dedicated to the presentation and discussion of our extensive, quantitative experimental evaluation of LamBERTa models.

[14] In both local and global settings, t-SNE was carried out until convergence, with the default value of *perplexity*, which is a key parameter to control the number of effective nearest neighbors in the manifold learning. Notably, we tried different values for the perplexity, both within and outside the recommended range (i.e., 5-50), but the difference in the visual analysis results between the global and local cases still remained the same.

Fig. 3.8 Explanations of LamBERTa-V1-NoDST (top) and LamBERTa-V2-NoDST (bottom) predictions for query *Who can be excluded from the testament?*

3.5.3 Explainability of LamBERTa models based on LIME

In this section, we take a different perspective, focusing on the interpretation of the models' prediction through the use of LIME - Local Interpretable Model-Agnostic Explanations [86]. LIME aims to explain the behavior of the underlying classifier "around" a query instance: by perturbing interpretable parts of the input query (i.e., words, for textual data), LIME weighs these perturbed data points by their proximity to the original query instance, and observes the associated predictions by the underlying classifier to determine which of those changes will have most impact on the prediction of the original query. To this purpose, the explanation is accomplished by approximating the underlying classifier locally by an interpretable one, such as a linear model. It should be noted that, even if the original classifier deals with non-linear complexities, like our LamBERTa models, in the neighborhood of an instance it behaves roughly as a linear model; therefore the local linear approximation can reasonably be assumed to be correct since LIME looks at a very small region around the query instance.

Figures 3.8–3.10 show LIME explanations of the LamBERTa models for various use-case queries. It should be noted that the queries have been defined in the form of questions and do not have strong syntactical patterns matching sentences of the corresponding relevant articles.

In line with the quantitative results discussed in the previous section, the token injection could be unnecessary to get the correct prediction. In this regard, one example is given by the query *Chi può essere escluso dal testamento?* (*Who can be excluded from the testament?*), shown in Figure 3.8, which contains all words appearing in the pre-trained Italian-BERT vocabulary. Indeed, the ground-truth article (Art. 463) is correctly predicted by LamBERTa-V2-NoDST

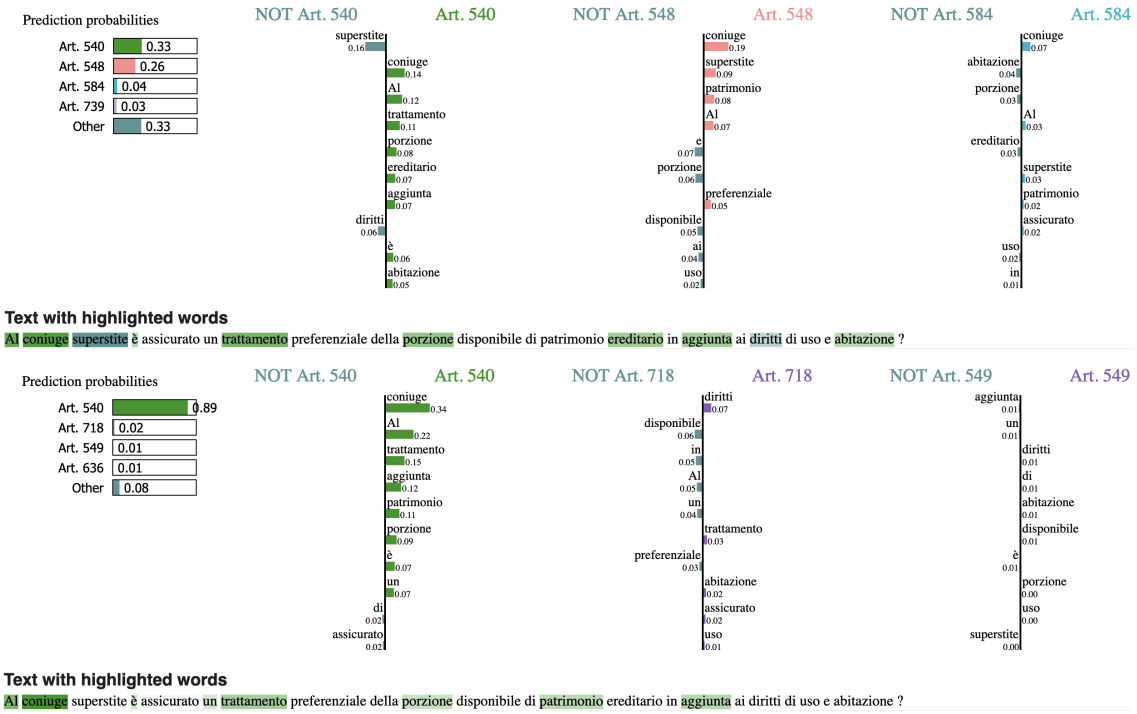


Fig. 3.9 Explanations of LamBERTa-V1 (top) and LamBERTa-V2 (bottom) predictions for query *Is the surviving spouse granted preferential treatment of the available portion of the hereditary patrimony in addition to the rights of use and habitation?*

and LamBERTa-V1-NoDST, where the latter leverage on more terms as it can be noticed by the LIME-explainer's highlighting on the important words.

Nonetheless, token injection can still be helpful in some cases. For instance, given query *Al coniuge superstite è assicurato un trattamento preferenziale della porzione disponibile di patrimonio ereditario in aggiunta ai diritti di uso e abitazione?* (*Is the surviving spouse granted preferential treatment of the available portion of the hereditary patrimony in addition to the rights of use and habitation?*), we find this is correctly answered by both LamBERTa-V1 and LamBERTa-V2 models that exploit ICC-specific token injection; for the sake of brevity, in Figure 3.9 we report explanations for models with random initialization of the injected tokens. This query contains terms 'superstite' and 'ereditario' that were added to the tokenizer, and hence are fully recognized by such models — otherwise they are subworded by those models equipped with the original pre-trained Italian-BERT vocabulary. Notably, although LamBERTa-V2 achieves higher prediction probability for the ground-truth article (Art. 540), LamBERTa-V1 behaves overall better as it is able to also retrieve the second most relevant article for the query (i.e., Art. 548). In this regard, we notice that while LamBERTa-V1's predictions are explained also by means of terms 'superstite' and 'ereditario', these are not recognized as essential by LamBERTa-V2.

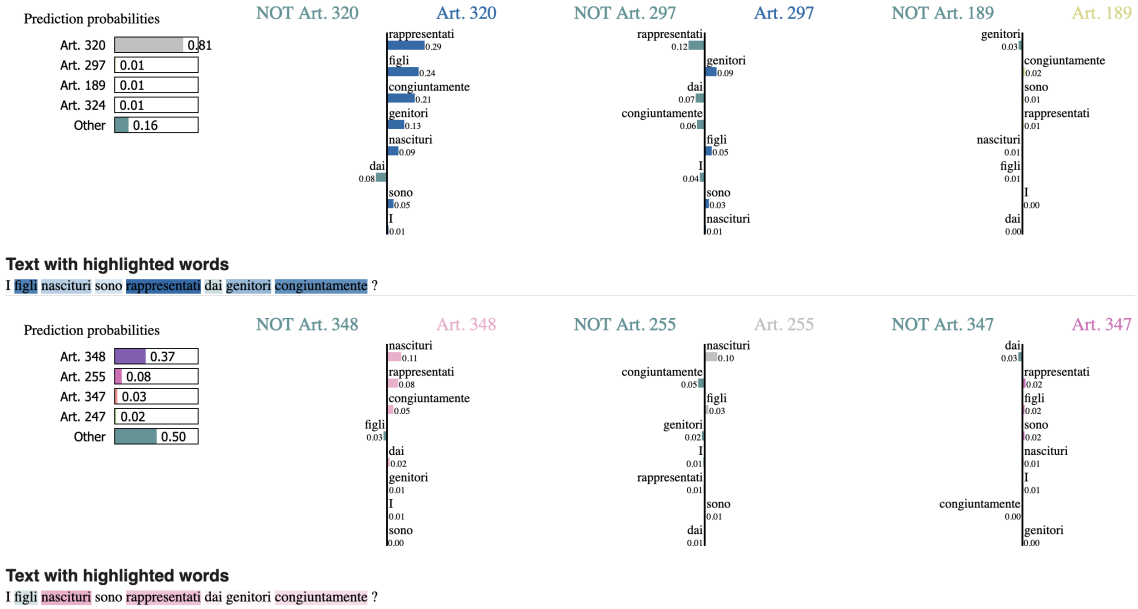


Fig. 3.10 Explanations of LamBERTa-V1 (top) and LamBERTa-V2-NoDST (bottom) predictions for query *Are the unborn children represented by the parents jointly?*

The benefit from using token injection is also evident for query *I figli nascituri sono rappresentati dai genitori congiuntamente?* (*Are the unborn children represented by the parents jointly?*), where 'nascituri' is missing from the pre-trained Italian-BERT vocabulary, and hence it is subworded (into 3 tokens). The ground-truth article (Art. 320) is strongly predicted by LamBERTa-V1 and LamBERTa-V1-ReDST, but is not by the counterpart LamBERTa-V2 models: looking at the LIME explanation (Figure 3.10), the injected token 'nascituri' is well-recognized in LamBERTa-V1 as an important feature for prediction along with other neighboring words in the query, but this does not hold for LamBERTa-V2. This may hint at a phenomenon of fresh-knowledge acquisition, by a model (i.e., LamBERTa-V1) that fine-tunes a general-domain one, against early-knowledge update exhibited by a model (i.e., LamBERTa-V2) that fine-tunes a domain-specific one.

Table 3.15 Article-driven, Clustering-based multi-label evaluation: comparison of results obtained by local models, with labeling scheme UniRR.T$^+$, based on clusters over TF-IDF representation vs. clusters over LamBERTa embeddings of the ICC articles, for all sets of book-sentence-queries (QType-1), paraphrased-sentence-queries (QType-2), comment-queries (QType-3), comment-sentence-queries (QType-4), and case-queries (QType-5). (Bold values correspond to the best model for each query-set and evaluation criterion)

query-set		R		P		F^{μ}		F^{M}	
	i	TF-IDF vectors	LamBERTa embeddings	TF-IDF vectors	LamBERTa embeddings	TF-IDF vectors	LamBERTa embeddings	TF-IDF vectors	LamBERTa embeddings
QType-1	Q_1	0.615	**0.641**	0.617	**0.643**	0.615	**0.641**	0.616	**0.642**
	Q_2	0.674	**0.682**	0.684	**0.687**	0.676	**0.684**	0.679	**0.685**
	Q_3	0.623	**0.677**	0.626	**0.674**	0.624	**0.675**	0.625	**0.675**
	Q_4	0.331	**0.454**	0.326	**0.456**	0.324	**0.454**	0.328	**0.455**
	Q_5	0.631	**0.719**	0.633	**0.722**	0.628	**0.719**	0.632	**0.720**
	Q_6	0.711	**0.781**	0.720	**0.784**	0.710	**0.782**	0.715	**0.782**
QType-2	Q_1	0.557	**0.575**	0.559	**0.577**	0.557	**0.575**	0.558	**0.576**
	Q_2	0.595	**0.662**	0.604	**0.666**	0.598	**0.663**	0.599	**0.664**
	Q_3	0.563	**0.859**	0.566	**0.884**	0.563	**0.849**	0.564	**0.871**
	Q_4	0.172	**0.748**	0.213	**0.769**	0.186	**0.727**	0.190	**0.759**
	Q_5	0.522	**0.606**	0.526	**0.609**	0.522	**0.606**	0.524	**0.607**
	Q_6	0.624	**0.682**	0.628	**0.690**	0.625	**0.684**	0.626	**0.686**
QType-3	Q_1	0.312	**0.329**	0.315	**0.333**	0.312	**0.331**	0.313	**0.331**
	Q_2	0.296	**0.320**	0.300	**0.322**	0.297	**0.320**	0.298	**0.321**
	Q_3	0.332	**0.349**	0.337	**0.353**	0.333	**0.350**	0.334	**0.351**
	Q_4	0.234	**0.249**	0.241	**0.253**	0.236	**0.250**	0.237	**0.251**
	Q_5	0.209	**0.330**	0.214	**0.333**	0.209	**0.330**	0.211	**0.331**
	Q_6	0.361	**0.372**	0.365	**0.373**	0.361	**0.371**	0.363	**0.372**
QType-4	Q_1	0.227	**0.230**	0.227	**0.234**	0.226	**0.231**	0.227	**0.232**
	Q_2	0.205	**0.222**	0.207	**0.223**	0.205	**0.223**	0.206	**0.223**
	Q_3	0.259	**0.265**	0.261	**0.268**	0.258	**0.265**	0.260	**0.266**
	Q_4	0.163	**0.171**	0.164	**0.175**	0.164	**0.172**	0.163	**0.173**
	Q_5	0.151	**0.180**	0.152	**0.183**	0.150	**0.180**	0.151	**0.181**
	Q_6	0.236	**0.246**	0.238	**0.250**	0.235	**0.247**	0.237	**0.248**
QType-5	Q_1	0.283	**0.290**	0.289	**0.294**	0.284	**0.292**	0.286	**0.292**
	Q_2	0.294	**0.300**	0.304	**0.311**	0.298	**0.305**	0.299	**0.274**
	Q_3	0.298	**0.393**	0.302	**0.391**	0.299	**0.392**	0.300	**0.392**
	Q_4	0.274	**0.282**	0.276	**0.285**	0.275	**0.282**	0.275	**0.283**
	Q_5	0.378	**0.407**	0.382	**0.411**	0.378	**0.408**	0.380	**0.409**
	Q_6	0.404	**0.454**	0.408	**0.456**	0.405	**0.454**	0.406	**0.455**

3.6 Appendix

Table 3.15 compares article-driven multi-label performance results based on clusters of articles modeled either through TF-IDF vectors or LamBERTa embeddings. Note that, to ease the comparison, the results under the 'TF-IDF vectors' columns correspond to the performance of local models reported in Table 3.4.

Table 3.16 Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-1	1: ‘persone fisiche’, 2: ‘persone giuridiche’, 3: ‘disposizioni generali’, 4: ‘associazioni e fondazioni’, 5: ‘associazioni non riconosciute e comitati’, 6: ‘domicilio e residenza’, 7: ‘assenza e dichiarazione di morte presunta’, 8: ‘assenza’, 9: ‘dichiarazione di morte presunta’, 10: ‘ragioni eventuali che competono alla persona di cui si ignora l’esistenza o di cui è stata dichiarata la morte presunta’, 11: ‘parentela e affinità’, 12: ‘matrimonio’, 13: ‘promessa di matrimonio’, 14: ‘matrimonio celebrato davanti a ministri del culto cattolico e matrimonio celebrato davanti a ministri dei culti ammessi nello stato’, 15: ‘matrimonio celebrato davanti all’ufficiale dello stato civile’, 16: ‘condizioni necessarie per contrarre matrimonio’, 17: ‘formalità preliminari del matrimonio’, 18: ‘opposizioni al matrimonio’, 19: ‘celebrazione del matrimonio’, 20: ‘matrimonio dei cittadini in paese straniero e degli stranieri nel regno’, 21: ‘nullità del matrimonio’, 22: ‘prove della celebrazione del matrimonio’, 23: ‘disposizioni penali’, 24: ‘diritti e doveri che nascono dal matrimonio’, 25: ‘scioglimento del matrimonio e separazione dei coniugi’, 26: ‘regime patrimoniale della famiglia’, 27: ‘fondo patrimoniale’, 28: ‘comunione legale’, 29: ‘comunione convenzionale’, 30: ‘regime di separazione dei beni’, 31: ‘impresa familiare’, 32: ‘stato di figlio’, 33: ‘presunzione di paternità’, 34: ‘prove della filiazione’, 35: ‘azione di disconoscimento e azioni di contestazione e di reclamo dello stato di figlio’, 36: ‘riconoscimento dei figli nati fuori dal matrimonio’, 37: ‘dichiarazione giudiziale della paternità e della maternità’, 38: ‘legittimazione dei figli naturali’, 39: ‘adozione di persone maggiori di età’, 40: ‘adozione di persone maggiori di età e suoi effetti’, 41: ‘forme dell’adozione di persone di maggiore età’, 42: ‘responsabilità genitoriale e diritti e doveri del figlio’, 43: ‘diritti e doveri del figlio’, 44: ‘esercizio della responsabilità genitoriale a seguito di separazione, scioglimento, cessazione degli effetti civili, annullamento, nullità del matrimonio ovvero all’esito di procedimenti relativi ai figli nati fuori del matrimonio’, 45: ‘ordini di protezione contro gli abusi familiari’, 46: ‘tutela e emancipazione’, 47: ‘tutela dei minori’, 48: ‘giudice tutelare’, 49: ‘tutore e protutore’, 50: ‘esercizio della tutela’, 51: ‘cessazione del tutore dall’ufficio’, 52: ‘rendimento del conto finale’, 53: ‘emancipazione’, 54: ‘affiliazione e affidamento’, 55: ‘misure di protezione delle persone prive in tutto od in parte di autonomia’, 56: ‘amministrazione di sostegno’, 57: ‘interdizione, inabilitazione e incapacità naturale’, 58: ‘alimenti’, 59: ‘atti dello stato civile’	1: ‘natural persons’, 2: ‘legal persons’, 3: ‘general provisions’, 4: ‘associations and foundations’, 5: ‘unrecognized associations and committees’, 6: ‘domicile and residence’, 7: ‘absence and declaration of presumed death’, 8: ‘absence’, 9: ‘declaration of presumed death’, 10: ‘possible reasons for the person whose existence is unknown or whose existence has been declared presumed death’, 11: ‘kinship and affinity’, 12: ‘marriage’, 13: ‘promise of marriage’, 14: ‘marriage celebrated before Catholic ministers and marriage celebrated before ministers of worship admitted in the state’, 15: ‘marriage celebrated before the registrar’, 16: ‘necessary conditions for marriage’, 17: ‘preliminary formalities for marriage’, 18: ‘opposition to marriage’, 19: ‘celebration of marriage’, 20: ‘marriage of citizens in a foreign country and foreigners in the kingdom’, 21: ‘nullity of marriage’, 22: ‘proof of celebration of marriage’, 23: ‘penal provisions’, 24: ‘rights and duties arising from marriage’, 25: ‘dissolution of marriage and separation of spouses’, 26: ‘family property regime’, 27: ‘patrimonial fund’, 28: ‘legal community’, 29: ‘conventional community’, 30: ‘property separation regime’, 31: ‘family business’, 32: ‘child status’, 33: ‘presumption of paternity’, 34: ‘evidence of filiation’, 35: ‘action of denial and actions of contestation and complaint of the child status’, 36: ‘recognition of children born out of wedlock’, 37: ‘judicial declaration of paternity and maternity’, 38: ‘legitimation of natural children’, 39: ‘adoption of older people’, 40: ‘adoption of older people and its effects’, 41: ‘forms of adoption of older people’, 42: ‘parental responsibility and rights and duties of the child’, 43: ‘rights and duties of the child’, 44: ‘practice of parental responsibility following separation, dissolution, termination of civil effects, annulment, nullity of marriage or the outcome of proceedings relating to children born out of wedlock’, 45: ‘protection orders against family abuse’, 46: ‘guardianship and emancipation’, 47: ‘guardianship of minors’, 48: ‘tutelary judge’, 49: ‘guardian and protutore’, 50: ‘practice of guardianship’, 51: ‘termination of the guardian from office’, 52: ‘return of the final account’, 53: ‘emancipation’, 54: ‘affiliation and custody’, 55: ‘protection measures for persons lacking in whole or in part autonomy’, 56: ‘support administration’, 57: ‘disqualification, disability and natural incapacity’, 58: ‘alimony’, 59: ‘civil status documents’

Table 3.17 *(cont.)* Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-2	1: ‘disposizioni generali sulle successioni’, 2: ‘apertura della successione, delazione e acquisto dell’eredità’, 3: ‘capacità di succedere’, 4: ‘indegnità’, 5: ‘rappresentazione’, 6: ‘accettazione dell’eredità’, 7: ‘disposizioni generali’, 8: ‘beneficio d’inventario’, 9: ‘separazione dei beni del defunto da quelli dell’erede’, 10: ‘rinunzia all’eredità’, 11: ‘eredità giacente’, 12: ‘petizione di eredità’, 13: ‘legittimari’, 14: ‘diritti riservati ai legittimari’, 15: ‘reintegrazione della quota riservata ai legittimari’, 16: ‘successioni legittime’, 17: ‘successione dei parenti’, 18: ‘successione del coniuge’, 19: ‘successione dello stato’, 20: ‘successioni testamentarie’, 21: ‘capacità di disporre per testamento’, 22: ‘capacità di ricevere per testamento’, 23: ‘forma dei testamenti’, 24: ‘testamenti ordinari’, 25: ‘testamenti speciali’, 26: ‘pubblicazione dei testamenti olografi e testamenti segreti’, 27: ‘istituzione di erede e legati’, 28: ‘disposizioni condizionali, a termine e modali’, 29: ‘legati’, 30: ‘diritto di accrescimento’, 31: ‘revocazione delle disposizioni testamentarie’, 32: ‘sostituzioni’, 33: ‘sostituzione ordinaria’, 34: ‘sostituzione fedecommissaria’, 35: ‘esecutori testamentari’, 36: ‘divisione’, 37: ‘collazione’, 38: ‘pagamento dei debiti’, 39: ‘effetti della divisione e garanzia delle quote’, 40: ‘annullamento e rescissione in materia di divisione’, 41: ‘patto di famiglia’, 42: ‘donazioni’, 43: ‘capacità di disporre e di ricevere per donazione’, 44: ‘forma e effetti della donazione’, 45: ‘revocazione delle donazioni’	1: ‘general provisions on succession’, 2: ‘opening of succession, delation and acquisition of inheritance’, 3: ‘ability to succeed’, 4: ‘unworthiness’, 5: ‘representation’, 6: ‘acceptance of the inheritance’, 7: ‘general provisions’, 8: ‘inventory benefit’, 9: ‘separation of the deceased’s assets from those of the heir’, 10: ‘disclaim of inheritance’, 11: ‘lying inheritance’, 12: ‘petition of inheritance’, 13: ‘heirs’, 14: ‘rights reserved for the heirs’, 15: ‘reinstatement of the quota reserved for the heirs’, 16: ‘legitimate succession’, 17: ‘succession of relatives’, 18: ‘succession of spouse’, 19: ‘succession of the state’, 20: ‘testamentary successions’, 21: ‘ability to dispose by testament’, 22: ‘ability to receive by testamemt’, 23: ‘form of testaments’, 24: ‘ordinary testaments ’, 25: ‘special testaments’, 26: ‘publication of holographic testaments and secret testaments’, 27: ‘establishment of heirs and legacies’, 28: ‘conditional provisions, forward and modal’, 29: ‘legacies’, 30: ‘right of accretion’, 31: ‘revocation of testamentary dispositions’, 32: ‘substitutions’, 33: ‘ordinary substitutions’, 34: ‘fidecommissary substitutions’, 35: ‘testamentary executors’, 36: ‘division’, 37: ‘collation’, 38: ‘payment of debts’, 39: ‘effects of division and guarantee of quotas’, 40: ‘annulment and rescission of division’, 41: ‘family pact’, 42: ‘donations’, 43: ‘ability to dispose and receive by donation’, 44: ‘forms and effects of donation’, 45: ‘revocation of donations’

Tables 3.16–3.21 list the ICC-heading attributes, for each of the ICC books, which were used in the task of attribute-aware law prediction.

Table 3.18 *(cont.)* Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-3	1: 'beni', 2: 'beni in generale', 3: 'beni nell'ordine corporativo', 4: 'beni immobili e mobili', 5: 'frutti', 6: 'beni appartenenti allo stato, agli enti pubblici e agli enti ecclesiastici', 7: 'proprietà', 8: 'disposizioni generali', 9: 'proprietà fondiaria', 10: 'riordinamento della proprietà rurale', 11: 'bonifica integrale', 12: 'vincoli idrogeologici e difese fluviali', 13: 'proprietà edilizia', 14: 'distanze nelle costruzioni, piantagioni e scavi, e muri, fossi e siepi interposti tra i fondi', 15: 'luci e vedute', 16: 'stillicidio', 17: 'acque', 18: 'modi di acquisto della proprietà', 19: 'occupazione e invenzione', 20: 'accessione, specificazione, unione e commistione', 21: 'azioni a difesa della proprietà', 22: 'superficie', 23: 'enfiteusi', 24: 'usufrutto, uso e abitazione', 25: 'usufrutto', 26: 'diritti nascenti dall'usufrutto', 27: 'obblighi nascenti dall'usufrutto', 28: 'estinzione e modificazioni dell'usufrutto', 29: 'uso e abitazione', 30: 'servitù prediali', 31: 'servitù coattive', 32: 'acquedotto e scarico coattivo', 33: 'appoggio e infissione di chiusa', 34: 'somministrazione coattiva di acqua a un edificio o a un fondo', 35: 'passaggio coattivo', 36: 'elettrodotto coattivo e passaggio coattivo di linee teleferiche', 37: 'servitù volontarie', 38: 'servitù acquistate per usucapione e per destinazione del padre di famiglia', 39: 'esercizio delle servitù', 40: 'estinzione delle servitù', 41: 'azioni a difesa delle servitù', 42: 'alcune servitù in materia di acque', 43: 'servitù di presa o di derivazione di acqua', 44: 'servitù degli scoli e degli avanzi di acqua', 45: 'comunione', 46: 'comunione in generale', 47: 'condominio negli edifici', 48: 'possesso', 49: 'effetti del possesso', 50: 'diritti e obblighi del possessore nella restituzione della cosa', 51: 'possesso di buona fede di beni mobili', 52: 'usucapione', 53: 'azioni a difesa del possesso', 54: 'denunzia di nuova opera e di danno temuto'	1: 'assets', 2: 'assets in general', 3: 'assets in the corporate order', 4: 'immovable and movable assets', 5: 'civil fruits', 6: 'assets belonging to the state, to public authorities and to ecclesiastical authorities', 7: 'property', 8: 'general provisions', 9: 'landed property', 10: 'rearrangement of rural property', 11: 'integral reclamation', 12: 'hydrogeological constraints and river defenses', 13: 'building property', 14: 'distances in buildings, plantations and excavations, and walls, ditches and hedges interposed between the bottoms', 15: 'lights and views', 16: 'dripping', 17: 'waters', 18: 'ways of purchasing property', 19 : 'occupation and invention', 20: 'accession, specification, union and admixture', 21: 'actions in defense of property', 22: 'surface', 23: 'emphyteusis', 24: 'usufruct, use and dwelling', 25: 'usufruct', 26: 'rights arising from usufruct', 27: 'obligations arising from usufruct', 28: 'extinction and modifications of usufruct', 29: 'use and dwelling', 30: 'predial servitude', 31: 'coercive servitude', 32: 'aqueduct and coercive drainage', 33: 'support and lock driving', 34: 'coercive administration of water to a building or ground', 35: 'coercive passage', 36: 'coercive power line and coercive passage of cableways', 37: 'voluntary servitude', 38: 'servitude purchased for usucapion and for the use of the father of a family', 39: 'practice of the servitude', 40: 'extinction of the servitude', 41:' actions in defense of the servitude', 42: 'some servitude concerning water', 43: 'servitude of water intake or diversion', 44: 'servitude of drains and leftovers of water', 45: ' communion', 46: 'communion in general', 47: 'condominium in buildings', 48: 'possession', 49: 'effects of possession', 50: 'rights and obligations of the owner to return the property', 51: 'possession of movable property in good faith', 52: 'usucaption', 53: 'actions in defense of possession', 54: 'denunciation of new work and feared damage'

Table 3.19 *(cont.)* Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-4	1: 'obbligazioni in generale', 2: 'disposizioni preliminari', 3: 'adempimento delle obbligazioni', 4: 'adempimento in generale', 5: 'pagamento con surrogazione', 6: 'mora del creditore', 7: 'inadempimento delle obbligazioni', 8: 'modi di estinzione delle obbligazioni diversi dall'adempimento', 9: 'novazione', 10: 'remissione', 11: 'compensazione', 12: 'confusione', 13: 'impossibilità sopravvenuta per causa non imputabile al debitore', 14: 'cessione dei crediti', 15: 'delegazione, espromissione e accollo', 16: 'alcune specie di obbligazioni', 17: 'obbligazioni pecuniarie', 18: 'obbligazioni alternative', 19: 'obbligazioni in solido', 20: 'obbligazioni divisibili e indivisibili', 21: 'contratti in generale', 22: 'requisiti del contratto', 23: 'accordo delle parti', 24: 'causa del contratto', 25: 'oggetto del contratto', 26: 'forma del contratto', 27: 'condizione nel contratto', 28: 'interpretazione del contratto', 29: 'effetti del contratto', 30: 'disposizioni generali', 31: 'clausola penale e caparra', 32: 'rappresentanza', 33: 'contratto per persona da nominare', 34: 'cessione del contratto', 35: 'contratto a favore di terzi', 36: 'simulazione', 37: 'nullità del contratto', 38: 'annullabilità del contratto', 39: 'incapacità', 40: 'vizi del consenso', 41: 'azione di annullamento', 42: 'rescissione del contratto', 43: 'risoluzione del contratto', 44: 'risoluzione per inadempimento', 45: 'impossibilità sopravvenuta', 46: 'eccessiva onerosità', 47: 'contratti del consumatore', 48: 'singoli contratti', 49: 'vendita', 50: 'obbligazioni del venditore', 51: 'obbligazioni del compratore', 52: 'riscatto convenzionale', 53: 'vendita di cose mobili', 54: 'vendita dei beni di consumo', 55: 'vendita con riserva di gradimento, a prova, a campione', 56: 'vendita con riserva della proprietà', 57: 'vendita su documenti e con pagamento contro documenti', 58: 'vendita a termine di titoli di credito', 59: 'vendita di cose immobili', 60: 'vendita di eredità', 61: 'riporto', 62: 'permuta', 63: 'contratto estimatorio', 64: 'somministrazione', 65: 'locazione', 66: 'locazione di fondi urbani', 67: 'affitto', 68: 'affitto di fondi rustici', 69: 'affitto a coltivatore diretto', 70: 'appalto', 71: 'trasporto', 72: 'trasporto di persone', 73: 'trasporto di cose', 74: 'mandato', 75: 'obbligazioni del mandatario', 76: 'obbligazioni del mandante', 77: 'estinzione del mandato', 78: 'commissione', 79: 'spedizione', 80: 'contratto di agenzia', 81: 'mediazione', 82: 'deposito', 83: 'deposito in generale', 84: 'deposito in albergo', 85: 'deposito nei magazzini generali', 86: 'sequestro convenzionale', 87: 'comodato', 88: 'mutuo', 89: 'conto corrente', 90: 'contratti bancari', 91: 'depositi bancari', 92: 'servizio bancario delle cassette di sicurezza', 93: 'apertura di credito bancario', 94: 'anticipazione bancaria', 95: 'operazioni bancarie in conto corrente', 96: 'sconto bancario', 97: 'rendita perpetua', 98: 'rendita vitalizia', 99: 'assicurazione', 100: 'assicurazione contro i danni', 101: 'assicurazione sulla vita', 102: 'riassicurazione', 103: 'disposizioni finali', 104: 'giuoco e scommessa', 105: 'fideiussione', 106: 'rapporti tra creditore e fideiussore', 107: 'rapporti tra fideiussore e debitore principale', 108: 'rapporti tra più fideiussori', 109: 'estinzione della fideiussione', 110: 'mandato di credito', 111: 'anticresi', 112: 'transazione', 113: 'cessione dei beni ai creditori', 114: 'promesse unilaterali', 115: 'titoli di credito', 116: 'titoli al portatore', 117: 'titoli all'ordine', 118: 'titoli nominativi', 119: 'gestione di affari', 120: 'pagamento dell'indebito', 121: 'arricchimento senza causa', 122: 'fatti illeciti'	1: 'obligations in general', 2: 'preliminary provisions', 3: 'fulfillment of obligations', 4: 'fulfillment in general', 5: 'payment with subrogation', 6: 'default of creditor', 7: 'default of obligations', 8: 'ways of extinguishing obligations other than performance', 9: 'novation', 10: 'remission', 11: 'set-off', 12: 'confusion', 13: 'impossibility occurred for reasons not attributable to the debtor', 14: 'assignment of credits', 15: 'delegation, expression and assumption', 16: 'some kinds of obligations', 17: 'pecuniary obligations', 18: 'alternative obligations', 19: 'joint and several obligations', 20: 'divisible and indivisible obligations', 21: 'contracts in general', 22: 'contract requirements', 23: 'agreement of the parties', 24 : 'cause of the contract', 25: 'object of the contract', 26: 'form of the contract', 27: 'condition in the contract', 28: 'interpretation of the contract', 29: 'effects of the contract', 30: 'general provisions', 31: 'penalty clause and deposit', 32: 'representation', 33: 'contract for person to be appointed', 34: 'transfer of the contract', 35: 'contract in favor of third parties', 36: 'simulation', 37: 'nullity of the contract', 38: 'voidability of the contract', 39: 'incapacity', 40: 'vices of consent', 41: 'action for annulment' , 42: 'termination of the contract', 43: 'resolution of the contract', 44: 'termination due to non-fulfillment', 45: 'occurred impossibility', 46: 'excessive onerousness', 47: 'consumer contracts', 48: 'individual contracts', 49: 'sale', 50: 'seller's obligations', 51: 'buyer's obligations', 52: 'conventional redemption', 53: 'sale of movable things', 54: 'sale of consumer goods', 55: 'sale subject to approval, trial, sample', 56: 'sale subject to ownership', 57: 'sale on documents and with payment against documents', 58: 'forward sale of debt securities', 59: 'sale of real estate', 60: 'sale of inheritance', 61: 'return', 62: 'exchange', 63: 'estimation contract', 64: 'administration', 65: 'lease', 66: 'lease of urban land', 67: 'rent', 68: 'rent of rustic land', 69: 'rent to direct farmer', 70: 'contract', 71: 'transport', 72: 'transport of people', 73: 'transport of things', 74: 'warrant', 75: 'agent's obligations', 76: 'principal's obligations', 77: 'termination of the warrant', 78: 'commission', 79: 'shipment', 80: 'agency contract', 81: 'mediation', 82: 'deposit', 83: 'deposit in general', 84: 'deposit in hotel', 85: 'deposit in general warehouses', 86: 'conventional confiscation', 87: 'free loan', 88: 'mortgage', 89: 'checking account', 90: 'bank contracts', 91: 'bank deposits', 92: 'safe deposit box banking service', 93: 'bank credit opening', 94: 'bank advance', 95: 'checking account banking', 96: 'bank discount', 97: 'perpetual income', 98: 'life annuity', 99: 'insurance', 100: 'damage insurance', 101: 'life insurance', 102: 'reinsurance', 103: 'final provisions', 104: 'game and bet', 105: 'surety', 106: 'relationship between creditor and guarantor', 107: 'relationship between guarantor and principal debtor', 108: 'relationship between several guarantors', 109: 'termination of surety', 110: 'credit warrant', 111: 'anticresi', 112: 'transaction', 113: 'transfer of assets to creditors', 114: 'unilateral promises', 115: 'debt securities', 116: 'bearer securities', 117: 'order securities', 118: 'registered securities', 119: 'business management', 120: 'undue payment', 121: 'causeless enrichment', 122: 'illegal deeds'

Table 3.20 *(cont.)* Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-5	1: 'disciplina delle attività professionali', 2: 'disposizioni generali', 3: 'ordinanze corporative e accordi economici collettivi', 4: 'contratto collettivo di lavoro e norme equiparate', 5: 'lavoro nell'impresa', 6: 'impresa in generale', 7: 'imprenditore', 8: 'collaboratori dell'imprenditore', 9: 'rapporto di lavoro', 10: 'costituzione del rapporto di lavoro', 11: 'diritti e obblighi delle parti', 12: 'previdenza e assistenza', 13: 'estinzione del rapporto di lavoro', 14: 'disposizioni finali', 15: 'tirocinio', 16: 'impresa agricola', 17: 'mezzadria', 18: 'colonia parziaria', 19: 'soccida', 20: 'soccida semplice', 21: 'soccida parziaria', 22: 'soccida con conferimento di pascolo', 23: 'disposizione finale', 24: 'imprese commerciali e altre imprese soggette a registrazione', 25: 'registro delle imprese', 26: 'obbligo di registrazione', 27: 'disposizioni particolari per le imprese commerciali', 28: 'rappresentanza', 29: 'scritture contabili', 30: 'insolvenza', 31: 'lavoro autonomo', 32: 'professioni intellettuali', 33: 'lavoro subordinato in particolari rapporti', 34: 'lavoro domestico', 35: 'società', 36: 'società semplice', 37: 'rapporti tra i soci', 38: 'rapporti con i terzi', 39: 'scioglimento della società', 40: 'scioglimento del rapporto sociale limitatamente a un socio', 41: 'società in nome collettivo', 42: 'società in accomandita semplice', 43: 'società per azioni', 44: 'costituzione per pubblica sottoscrizione', 45: 'promotori e soci fondatori', 46: 'patti parasociali', 47: 'conferimenti', 48: 'azioni e altri strumenti finanziari partecipativi', 49: 'assemblea', 50: 'amministrazione e controllo', 51: 'amministratori', 52: 'collegio sindacale', 53: 'revisione legale dei conti', 54: 'sistema dualistico', 55: 'sistema monistico', 56: 'obbligazioni', 57: 'libri sociali', 58: 'bilancio', 59: 'modificazioni dello statuto', 60: 'patrimoni destinati ad uno specifico affare', 61: 'società con partecipazione dello stato o di enti pubblici', 62: 'società di interesse nazionale', 63: 'società in accomandita per azioni', 64: 'società a responsabilità limitata', 65: 'conferimenti e quote', 66: 'amministrazione della società e controlli', 67: 'decisioni dei soci', 68: 'modificazioni dell'atto costitutivo', 69: 'scioglimento e liquidazione delle società di capitali', 70: 'direzione e coordinamento di società', 71: 'trasformazione, fusione e scissione', 72: 'trasformazione', 73: 'fusione delle società', 74: 'scissione delle società', 75: 'società costituite all'estero', 76: 'società cooperative e mutue assicuratrici', 77: 'società cooperative', 78: 'disposizioni generali di società cooperative a mutualità prevalente', 79: 'costituzione', 80: 'quote e azioni', 81: 'organi sociali', 82: 'controlli', 83: 'mutue assicuratrici', 84: 'associazione in partecipazione', 85: 'azienda', 86: 'ditta e insegna', 87: 'marchio', 88: 'diritti sulle opere dell'ingegno e sulle invenzioni industriali', 89: 'diritto di autore sulle opere dell'ingegno letterarie e artistiche', 90: 'diritto di brevetto per invenzioni industriali', 91: 'diritto di brevetto per modelli di utilità e di registrazione per disegni e modelli', 92: 'disciplina della concorrenza e dei consorzi', 93: 'disciplina della concorrenza', 94: 'concorrenza sleale', 95: 'consorzi per il coordinamento della produzione e degli scambi', 96: 'consorzi con attività esterna', 97: 'consorzi obbligatori', 98: 'controlli dell'autorità governativa', 99: 'disposizioni penali in materia di società, di consorzi e di altri enti privati', 100: 'falsità', 101: 'illeciti commessi dagli amministratori', 102: 'illeciti commessi mediante omissione', 103: 'altri illeciti, circostanze attenuanti e misure di sicurezza patrimoniali'	1: 'regulation of professional activities', 2: 'general provisions', 3: 'corporate regulations and collective economic agreements', 4: 'collective bargaining agreement and equivalent rules', 5: 'work in enterprise', 6: 'enterprise in general', 7: 'entrepreneur', 8: 'collaborators of the entrepreneur', 9: 'employment relationship', 10: 'establishment of the employment relationship', 11: 'rights and obligations of the parties', 12: 'social security and assistance', 13: 'termination of the employment relationship', 14: 'final provisions', 15: 'apprenticeship', 16: 'agricultural enterprise', 17: 'sharecropping', 18: 'partial colony', 19: 'soccida', 20: 'simple soccida', 21: 'partial soccida', 22: 'soccida with granting of pasture', 23: 'final provision', 24: 'commercial enterprises and other enterprises subject to registration', 25: 'business register', 26: 'obligation of registration', 27: 'special provisions for commercial enterprises', 28: 'representation', 29: 'accounting records', 30: 'insolvency', 31: 'self-employment', 32: 'intellectual professions', 33: 'subordinate work in particular relationships', 34: 'domestic work', 35: 'company', 36: 'simple company', 37: 'relationships between shareholders', 38: 'relationships with third parties', 39: 'dissolution of the company', 40: 'dissolution of the social relationship limited to one shareholder', 41: 'collective denomination company', 42: 'limited partnership company', 43: 'joint stock company', 44: 'constitution by public subscription', 45: 'promoters and founding partners', 46: 'shareholders' agreements', 47: 'contributions', 48: 'shares and other equity financial instruments', 49: 'shareholders' meeting', 50: 'administration and control', 51: 'directors', 52: 'board of statutory auditors', 53: 'statutory audit', 54: 'two-tier system', 55: 'one-tier system', 56: 'bonds', 57: 'company books', 58: 'balance sheet', 59: 'amendments to the statute', 60: 'assets intended for a specific business', 61: 'company with state participation or public authorities', 62: 'company of national interest', 63: 'company limited by shares', 64: 'company with limited responsibility', 65: 'contributions and shares', 66: 'company administration and controls', 67: 'shareholders' decisions', 68: 'amendments to the articles of association', 69: 'dissolution and liquidation of corporations', 70: 'management and coordination of companies', 71: 'transformation, merger and spin-off', 72: 'transformation', 73: 'merger of companies', 74: 'spin-off of companies', 75: 'companies incorporated abroad', 76: 'cooperative companies and mutual insurance companies', 77: 'cooperative companies', 78:' general provisions of cooperative societies with prevalent mutuality', 79: 'incorporation', 80: 'quotas and shares', 81: 'corporate bodies', 82: 'controls', 83: 'mutual insurance companies', 84: 'association in participation', 85: 'firm', 86: 'firm and sign', 87: 'trademark', 88: 'intellectual property rights and industrial inventions', 89: 'copyright on literary and artistic intellectual works', 90: 'patent law for industrial inventions', 91: 'patent law for utility models and registration for designs and models', 92: 'competition and consortium regulations', 93: 'competition rules', 94: 'unfair competition', 95: 'consortia for the coordination of production and trade', 96: 'consortia with external activities', 97: 'compulsory consortia', 98: 'controls by the governmental authority', 99: 'criminal provisions concerning companies, consortia and other private authorities', 100: 'falsity', 101: 'offenses committed by administrators', 102: 'offenses committed by omission', 103: 'other offenses, mitigating circumstances and asset security measures'

Table 3.21 *(cont.)* Unique headings of book subdivisions originally provided in the ICC and used in the attribute-aware article prediction evaluation task

	attributes	attributes (translated in English)
Book-6	1: 'trascrizione', 2: 'trascrizione degli atti relativi ai beni immobili', 3: 'pubblicità dei registri immobiliari e responsabilità dei conservatori', 4: 'trascrizione degli atti relativi ad alcuni beni mobili', 5: 'trascrizione relativamente alle navi, agli aeromobili e agli autoveicoli', 6: 'trascrizione relativamente ad altri beni mobili', 7: 'prove', 8: 'disposizioni generali', 9: 'prova documentale', 10: 'atto pubblico', 11: 'scrittura privata', 12: 'scritture contabili delle imprese soggette a registrazione', 13: 'riproduzioni meccaniche', 14: 'taglie o tacche di contrassegno', 15: 'copie degli atti', 16: 'atti di ricognizione o di rinnovazione', 17: 'prova testimoniale', 18: 'presunzioni', 19: 'confessione', 20: 'giuramento', 21: 'responsabilità patrimoniale, cause di prelazione e conservazione della garanzia patrimoniale', 22: 'privilegi', 23: 'privilegi sui mobili', 24: 'privilegi generali sui mobili', 25: 'privilegi sopra determinati mobili', 26: 'privilegi sopra gli immobili', 27: 'ordine dei privilegi', 28: 'pegno', 29: 'pegno dei beni mobili', 30: 'pegno di crediti e altri diritti', 31: 'ipoteche', 32: 'ipoteca legale', 33: 'ipoteca giudiziale', 34: 'ipoteca volontaria', 35: 'iscrizione e rinnovazione delle ipoteche', 36: 'iscrizione', 37: 'rinnovazione', 38: 'ordine delle ipoteche', 39: 'effetti dell'ipoteca rispetto al terzo acquirente', 40: 'effetti dell'ipoteca rispetto al terzo datore', 41: 'riduzione delle ipoteche', 42: 'estinzione delle ipoteche', 43: 'cancellazione dell'iscrizione', 44: 'modo di liberare i beni dalle ipoteche', 45: 'rinunzia e astensione del creditore nell'espropriazione forzata', 46: 'mezzi di conservazione della garanzia patrimoniale', 47: 'azione surrogatoria', 48: 'azione revocatoria', 49: 'sequestro conservativo', 50: 'tutela giurisdizionale dei diritti', 51: 'esecuzione forzata', 52: 'espropriazione', 53: 'effetti del pignoramento', 54: 'effetti della vendita forzata e dell'assegnazione', 55: 'espropriazione di beni oggetto di vincoli di indisponibilità o di alienazioni a titolo gratuito', 56: 'esecuzione forzata in forma specifica', 57: 'prescrizione e decadenza', 58: 'prescrizione', 59: 'sospensione della prescrizione', 60: 'interruzione della prescrizione', 61: 'termine della prescrizione', 62: 'prescrizione ordinaria', 63: 'prescrizioni brevi', 64: 'prescrizioni presuntive', 65: 'computo dei termini', 66: 'decadenza'	1: 'transcription', 2: 'transcription of deeds relating to immovable properties', 3: 'advertising of real estate registers and liability of conservators', 4: 'transcription of deeds relating to some movable property', 5: 'transcription relating to ships, aircraft and motor vehicles', 6: 'transcription relating to other movable properties', 7: 'evidence', 8: 'general provisions', 9: 'documentary evidence', 10: 'public deed', 11: 'private deed', 12: 'accounting records of companies subject to registration', 13: 'mechanical reproductions', 14: 'sizes or mark notches', 15: 'copies of deeds', 16: 'acts of recognition or renewal', 17: 'witness evidence', 18: 'presumptions', 19: 'confession', 20: 'oath', 21: 'property liability, causes of preemption and retention of the patrimonial guarantee', 22: 'privileges', 23: 'privileges on movables', 24: 'general privileges on movables', 25: 'privileges over certain movables', 26: 'privileges over real estate', 27: 'order of privileges', 28: 'pledge', 29: 'pledge of movable property', 30: 'pledge of credits and other rights', 31: 'mortgages', 32: 'legal mortgage', 33: 'judicial mortgage', 34: 'voluntary mortgage', 35: 'registration and renewal of mortgages', 36: 'registration', 37: 'renewal', 38: 'mortgage order', 39: 'effects of 'mortgage with respect to the third buyer', 40: 'effects of the mortgage with respect to the third employer', 41: 'reduction of mortgages', 42: 'repayment of mortgages', 43: 'cancellation of registration', 44: 'way to release the assets from mortgages', 45: 'renunciation and abstention of the creditor in forced expropriation', 46: 'means of preserving the patrimonial guarantee', 47: 'subrogation action', 48: 'revocatory action', 49: 'conservative seizure', 50: 'judicial protection of rights', 51: 'forced execution ', 52: 'expropriation', 53: 'effects of foreclosure', 54: 'effects of forced sale and assignment', 55: 'expropriation of assets subject to restrictions of unavailability or free disposal', 56: 'specific forced execution', 57: 'prescription and forfeiture', 58: 'prescription', 59: 'suspension of prescription', 60: 'termination of prescription', 61: 'limitation period of prescription', 62: 'ordinary prescription', 63: 'short prescription', 64: 'presumptive prescription', 65: 'calculation of time limits', 66: 'forfeiture'

Chapter 4

LamBERTa applied on the GDPR

This chapter is based on the research work presented in the paper "GDPR Article Retrieval based on Domain-adaptive and Task-adaptive Legal Pre-trained Language Models" [91]. In that case, the LamBERTa models are applied on the before-mentioned GDPR corpus. To the best of our knowledge this is the first work to embrace the more general task of GDPR article retrieval by leveraging PLMs, and also powering them through self-supervised task-adaptive pre-training schemes. Firstly, in section 4.1 inspired by the state of art, we present (RSP) Related Sentence Prediction and (MCA) Mulitple choice Answering two task-adaptive pre-training of BERT models tailored to the GDPR article retrieval task . Following this, we provide a set of training models and evaluation methodologies, in section 3.2. Finally, in section 4.3 we present the Results derived from our experiments.

4.1 Self-supervised task-adaptive pre-training strategies

Task-adaptive fine-tuning is known to be the commonly used approach to enable a direct application of a pre-trained model to a downstream task. On the other hand, *domain-adaptive pre-training* allows for a more extensive customization of a pre-existing out-of-the-box model to a specialized language domain. In the legal domain, two approaches to domain-adaptive pre-training have been adopted: one is to continue pre-training the model using a legal corpus, while the other is to start the pre-training process from scratch using a legal corpus. An exemplary study adopting both approaches is provided in [20] for the development of the well-known Legal-BERT models.

However, the availability of legal data for a particular task may be limited, which can hinder the effective training of the model. Consequently, the model may struggle to fully grasp the meaning of legal texts and generalize the acquired knowledge in enough detail to handle unknown inputs successfully. It has been demonstrated that pre-training a model

on a legal corpus does not always guarantee significant improvements over fine-tuning a corresponding domain-general pre-trained model on the target task (e.g., [90]). According to [103, 37], the advantages of domain-specific pre-training are particularly evident when dealing with low-resource downstream tasks.

Nonetheless, there exists another form of pre-training, which is to train a (pre-trained) model on a smaller corpus of the specialized domain such that the corpus is directly related to the target task but its documents are not annotated with the target class labels. This form of unsupervised pre-training, called *task-adaptive pre-training*, has shown competitiveness in comparison to domain-adaptive pre-training and can also enhance performance when combined with it for the downstream task [37]. However, these findings have not been proven specifically for the legal domain, presenting opportunities for further research in this area.

In this work, we aim to fill the above gap by developing the first approaches to task-adaptive pre-training of BERT models tailored to the GDPR article retrieval task. We shall describe our two proposed approaches in the following sections.

4.1.1 Related Sentence Prediction

Besides Masked Language Modeling, BERT was unsupervisedly trained on another pre-training objective, called Next Sentence Prediction (NSP), to cover a variety of downstream tasks involving sentence pairs. Given two word sequences as input, the NSP task is to determine if the second sequence is subsequent to the first in a document.

Inspired by the idea underlying NSP, we propose the *Related Sentence Prediction* (RSP) approach to task-adaptive pre-training. Basically, RSP is to predict if two given sentences in input are related to each other or not; the *relatedness* concept can be defined in more ways, and in this work we shall consider the location of two sentences within the same GDPR article.

Let us denote with $\mathscr{A}$ and $\mathscr{R}$ the sets of GDPR articles and recitals, respectively. Each article A_i can be modeled as a sequence $A_i = (p_{i,1}, \ldots, p_{i,n_i})$, where ps denote textual units belonging to A_i, i.e., sentences – following an analogy with NSP – or, more generally, paragraphs constituting A_i. For any $A_i \in \mathscr{A}$, we define $\mathsf{RSP}_B(A_i)$, with $B \in \{\mathscr{A}, \mathscr{R}\}$, as a meta-function expressing *relatedness* between A_i and different portions of the GDPR, thus producing a set of training instances as associations between textual units of A_i and textual units from the document collection B. This means that, depending on the choice of B, sentences/paragraphs of A_i are coupled with either sentences/paragraphs from other article(s) than A_i, or with recitals. We refer to the first case (i.e., $B = \mathscr{A}$) as *article-level RSP*, and to the second case (i.e., $B = \mathscr{R}$) as *recital-level RSP*. In both cases, two types of training instances are built, which are *contrastive* to each other: the ones referring to "positive" relatedness

(denoted with superscript $+$) and the other ones referring to "negative" relatedness (denoted with superscript $-$). We shall provide our definitions next.

Article-level RSP. For any article $A_i \in \mathscr{A}$, we define

$$\mathsf{RSP}_{\mathscr{A}}(A_i) = \mathsf{RSP}^{+}(A_i, A_i) \cup \mathsf{RSP}^{-}(A_i, \mathscr{A} \setminus A_i), \tag{4.1}$$

where $\mathsf{RSP}^{+}(A_i, A_i)$ is a function producing a set of intra-article pairings over the textual units of A_i as positive relatedness associations, and $\mathsf{RSP}^{-}(A_i, \mathscr{A} \setminus A_i)$ is a function producing a set of inter-article pairings between textual units of A_i and textual units from articles different from A_i, as negative relatedness associations. The above functions are specified so as to satisfy the following minimum requirements. First, to ensure balance between positive and negative associations, the number of training instances as pairs derived from A_i, which is bounded by $|A_i|(|A_i|-1)/2$, is used to constrain the number of training instances derived by pairing units from A_i and units from any other article A_j, with $j \neq i$. Second, the choice of A_j is, by default, made uniformly at random, although constraints could be added to get A_j more or less "topically distant" from A_i (e.g., selecting A_j from the same chapter of A_i or from a different one). Third, multiple choices of A_j are made if $|A_j| < |A_i|$ (i.e., multiple articles need to be involved to form a number of negative associations to equal the number of positive ones); in general, using multiple articles against A_i can be useful to diversify the negative associations with A_i, thus obtaining a mix of negative training instances at different hardness levels.

Recital-level RSP. Leveraging recitals for training a model on an RSP task is strongly justified since they are originally conceived in the GDPR as essential complements for the articles. Based upon this, we provide a function definition analogous to the article-level one, which is as follows:

$$\mathsf{RSP}_{\mathscr{R}}(A_i) = \mathsf{RSP}^{+}(A_i, R^{(i)}) \cup \mathsf{RSP}^{-}(A_i, \mathscr{R} \setminus R^{(i)}), \tag{4.2}$$

where $R^{(i)}$ denotes the set of recitals associated with article A_i, the positive relatedness function $\mathsf{RSP}^{+}(A_i, R^{(i)})$ yields a set of training instances as pairs obtained by the textual units of A_i and the recitals in $R^{(i)}$, and the negative relatedness function $\mathsf{RSP}^{-}(A_i, \mathscr{R} \setminus R^{(i)})$ yields a set of training instances as pairs obtained by textual units of A_i and recitals not in $R^{(i)}$. It should be noted that we specify associations between portions of articles and the entire recitals, as we want to provide a maximal context based on recitals for each of the articles.

4.1.2 Multiple Choice Answering

Our second proposed task-adaptive pre-training approach is *Multiple Choice Answering* (MCA), which is defined as choosing the correct answer from a set of possible answers relating to an input query. In a sense, this task can be seen as a blend between the RSP task and the Masked Language Modeling task, which is at the core of BERT-like pre-training. In fact, the latter requires to choose the correct word to fill in a mask from a set of possible options based on the context of an input sentence. Analogously, MCA requires to choose from a set of sentences that are presented as either positively or negatively related to an input one, in a similar fashion to the RSP task. Also, we again consider the opportunity of enriching the training through the recitals, therefore we shall distinguish between *article-level MCA* and *recital-level MCA*.

Article-level MCA. Given an integer $k > 1$ as the number of choices, for any article $A_i \in \mathscr{A}$, we define

$$\mathsf{MCA}_{\mathscr{A},k}(A_i) = \bigcup_{j=1}^{|A_i|} \langle \mathsf{MCA}^{+}(p_{i,j}, A_i), \mathsf{MCA}^{-}(p_{i,j}, \mathscr{A} \setminus A_i) \rangle, \tag{4.3}$$

where $\mathsf{MCA}^{+}(p_{i,j}, A_i)$ yields a pairing between $p_{i,j}$ and another unit randomly chosen from A_i (i.e., the positive or correct choice for $p_{i,j}$), and $\mathsf{MCA}^{-}(p_{i,j}, \mathscr{A} \setminus A_i)$ yields a $(k-1)$-sized set of pairings between $p_{i,j}$ and units each selected from randomly chosen articles A_j, with $j \neq i$. Overall, for each A_i, a set of training instances is computed, where each training instance is a tuple of size $k+1$ (i.e., a sentence/paragraph and its relating k choices). Note that the positions of the k choices are actually randomly shuffled so that the position of the correct choice is variable through all training instances.

Recital-level MCA. By changing the answering choice context from articles to recitals, we have the following definition:

$$\mathsf{MCA}_{\mathscr{R},k}(R^{(i)}) = \bigcup_{j=1}^{|A_i|} \langle \mathsf{MCA}^{+}(p_{i,j}, R^{(i)}), \mathsf{MCA}^{-}(p_{i,j}, \mathscr{R} \setminus R^{(i)}) \rangle, \tag{4.4}$$

where $R^{(i)}$ denotes the set of recitals associated with article A_i, and the positive, resp. negative, MCA functions have analogous definitions to the corresponding article-level ones. Note however that, consistently with the recital-level RSP definition, recitals are considered as atomic units.

Table 4.1 Summary of our developed models fine-tuned on the GDPR article retrieval task

Model name	task-adaptive pre-training	data-enrichment	Model name	task-adaptive pre-training	data-enrichment
BERT	✗	✗	BERT-r	✗	✓
LegalBERT	✗	✗	LegalBERT-r	✗	✓
EULegalBERT	✗	✗	EULegalBERT-r	✗	✓
BERT-RSP	RSP	✗	BERT-RSP-r	RSP	✓
BERT-MCA	MCA	✗	BERT-MCA-r	MCA	✓
LegalBERT-RSP	RSP	✗	LegalBERT-RSP-r	RSP	✓
LegalBERT-MCA	MCA	✗	LegalBERT-MCA-r	MCA	✓
EULegalBERT-RSP	RSP	✗	EULegalBERT-RSP-r	RSP	✓
EULegalBERT-MCA	MCA	✗	EULegalBERT-MCA-r	MCA	✓

4.2 Training and evaluation methodologies

4.2.1 Model selection and settings

Our study is versatile w.r.t. the choice of PLM to deal with the GDPR search and retrieval. As happened in the past for other novel applications of PLMs, demonstration through BERT (`bert-base-uncased`) [25] represents a primary choice. Moreover, this allows us to naturally couple evaluation based on a domain-general BERT model with evaluation based on legal specialized counterparts, which are the well-known family of models in [20]. Specifically, we use (i) the main model, named LegalBERT (`legal-bert-base-uncased`), which was pre-trained from scratch on large corpora including EU legislation, US contracts and cases, and UK legislation; (ii) a EU specific model, named EULegalBERT (`bert-base-uncased-eurlex`), which was further pre-trained starting from BERT base using EU legislation only.[1]

Table 4.1 provides a summary of our developed models. Suffix -r is used to denote the *recital-enriched* models, i.e., the recital-level RSP or MCA based models, as well as the base BERT, LegalBERT, and EULegalBERT which were fine-tuned based on articles and recitals as the training data. Note also that, in the fine-tuning stage, each of the models was trained for 10 epochs, using cross-entropy as loss function, AdamW optimizer and initial learning rate selected within [1e-5, 5e-5] on batches of 256 examples.

4.2.2 Training data preparation

Data enhancement. GDPR articles exhibit three distinctive traits in their logical structure. First, the text of an article is commonly organized as a numbered sequence of paragraphs (commas), where a paragraph can sometimes include an enumerated list of points. Second,

[1] LEGAL-BERT models are available at https://huggingface.co/nlpaueb/legal-bert-base-uncased

an article often contains references to specific paragraphs or points of the same article as well as of other articles. Third, one or more recitals can be associated with specific paragraphs, subparagraphs, or points within the same article.

Such features of the GDPR articles prompted us to carry out some preprocessing aimed to enhance them for feeding a language model. In particular, we pursued a threefold goal: (i) to enrich the article contents by expanding the references to (portions of) other articles or chapters; (ii) to refactor structured parts of an article in order to resolve anaphoric passages; and (iii) to produce a recital-based labeling of the articles at the finest level, by leveraging associations of recitals to the individual paragraphs of an article, when available. While the latter required no particular effort, the first two objectives were accomplished through a semi-automatic process with manual supervision to produce a reliable outcome. Specifically, with regard to objective (i), we resolved each reference by replacing it either with the original text of the referred part or with an inferred short description, in the form of a citation (i.e., enclosed by quotation marks). Concerning objective (ii), any enumerated list of points in an article was either replaced by as many paragraphs as the number of points, each equally preceded by the common premise of the point list, or just flattened by keeping the premise once followed by the enumeration, in case of relatively short texts as points in the list.

Fine-tuning article labeling schemes. Creating a training dataset for the downstream task, i.e., GDPR article retrieval, requires that the entire corpus must be used to embed its knowledge fully, and each class label must correspond to a specific GDPR article since we want to learn how to classify at the article level. To this purpose, in order to build the training set for the fine-tuning task, we resort to an unsupervised article-labeling scheme proposed before in section 4.1 “Methods for unsupervised training-instance labeling” which is designed to select and combine portions of each article, while ensuring balance of the contributions of each article, which clearly have different lengths. This scheme applies a round-robin method to iterate over replicas of the same group of training instances per article until a minimum number of instances (to be produced for each article) is reached. More specifically, the *unigram with parameterized emphasis on the title* scheme creates a set of training instances for each article which is comprised of round-robin selected sentences from the article, along with replicas of the article’s title.

4.2.3 Query sets

We are not aware of any publicly available benchmark for evaluating retrieval models on GDPR data. Therefore, we built our own query data as test sets, by varying them in terms of source, length, and lexical characteristics. We define the following query sets: **QA**, which

contains sentences randomly extracted from the GDPR articles; **QpA**, where each sentence in QA is *paraphrased* through an English-Spanish-Italian-English machine translation of the queries (via Google Translate); **QR** and **QpR**, which are analogous of QA and QpA, respectively, but replacing articles with recitals; **QC**, which contains *expert commentary* texts related to the GDPR articles, i.e., a set of opinions and comments provided by experts in data protection and privacy; **QCs**, where each comment in QC is broken down into its constituting sentences.

Each of the **QA** and **QpA**, resp. **QR** and **QpR**, sets contains 661, resp. 138, queries, with an average of 60 ($\pm$ 35), resp. 112 ($\pm$ 61), words per query. Also, **QC**, resp. **QCs**, contains 45, resp. 272, queries, with an average of 169 ($\pm$ 52), resp. 28 ($\pm$ 13), words per query.

4.2.4 Assessment criteria

Each query is associated with one article (ground-truth). For each article A_i, we first computed the following statistics: the recall for A_i (Rec_i), i.e., the number of queries A_i was correctly predicted out of all queries actually pertinent to A_i, the precision for A_i ($Prec_i$), i.e., the number of queries A_i was correctly predicted out of all predictions of A_i, and the F-measure for A_i, i.e., $F_i = 2Prec_iRec_i/(Prec_i + Rec_i)$. Then, we averaged over all articles to obtain the per-article average *precision* (*Prec*), *recall* (*Rec*), *micro-averaged F-measure* (F^μ) as the average over all F_is, and *macro-averaged F-measure* (F^M) as the harmonic mean of *Prec* and *Rec*.

In addition, we accounted for the top-3 predictions and the position (rank) of the correct article in predictions: the former is the fraction of correct article labels that are found in the top-3 predictions (i.e., top-3-probability results in response to each query), and averaging over all queries, which is the *recall*@3 (*Rec*@3); the latter is the *mean reciprocal rank* (*MRR*) considering for each query the rank of the correct prediction over the classification probability distribution, and averaging over all queries.

4.3 Results

We organize our presentation of the results into two parts: the first reporting the performance of our models trained on articles only, and the second considering the recital-enriched models.

4.3.1 Training on articles only

Comparison of BERT models. Table 4.2 shows results obtained by the base BERT, LegalBERT, and EULegalBERT on the various query sets.

Table 4.2 Performance results of base BERT, LegalBERT, and EULegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	*Rec*	*Prec*	F^{μ}	F^{M}	*Rec*@3	*MRR*
QA	BERT	0.992	0.982	0.985	0.987	0.992	0.988
	LegalBERT	**0.996**	**0.987**	**0.990**	**0.991**	**0.997**	**0.993**
	EULegalBERT	0.984	0.934	0.946	0.958	0.965	0.962
QpA	BERT	0.949	0.930	0.929	0.939	0.952	0.933
	LegalBERT	**0.975**	**0.944**	**0.951**	**0.959**	**0.971**	**0.962**
	EULegalBERT	0.878	0.793	0.812	0.833	0.900	0.870
QR	BERT	0.667	0.620	0.596	0.643	0.739	0.683
	LegalBERT	**0.703**	**0.674**	**0.658**	**0.688**	**0.804**	**0.740**
	EULegalBERT	0.519	0.483	0.451	0.501	0.609	0.557
QpR	BERT	0.624	0.574	0.550	0.598	0.717	0.657
	LegalBERT	**0.707**	**0.685**	**0.662**	**0.696**	**0.775**	**0.726**
	EULegalBERT	0.371	0.325	0.291	0.347	0.558	0.478
QCs	BERT	0.340	0.590	0.390	0.431	**0.570**	0.519
	LegalBERT	**0.357**	**0.668**	**0.419**	**0.465**	0.548	**0.482**
	EULegalBERT	0.215	0.425	0.262	0.285	0.349	0.327
QC	BERT	0.556	0.669	0.590	0.607	0.778	0.752
	LegalBERT	**0.794**	**0.862**	**0.806**	**0.826**	**0.889**	**0.848**
	EULegalBERT	0.385	0.474	0.399	0.425	0.689	0.558

Table 4.3 Performance results of task-adaptive pre-trained BERT models and comparison with base BERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	*Rec*	*Prec*	F^{μ}	F^{M}	*Rec*@3	*MRR*
QA	BERT	0.992	0.982	0.985	0.987	0.992	0.988
	BERT-RSP	0.990	0.980	0.981	0.985	0.994	0.989
	BERT-MCA	**0.998**	**0.990**	**0.993**	**0.994**	**0.997**	**0.996**
QpA	BERT	**0.949**	**0.930**	**0.929**	**0.939**	0.952	0.933
	BERT-RSP	0.930	0.895	0.895	0.913	0.926	0.906
	BERT-MCA	0.943	0.920	0.917	0.931	**0.961**	**0.945**
QR	BERT	0.667	0.620	0.596	0.643	0.739	0.683
	BERT-RSP	**0.684**	0.629	0.606	0.655	0.710	0.665
	BERT-MCA	0.675	**0.644**	**0.615**	**0.659**	**0.768**	**0.701**
QpR	BERT	0.624	0.574	0.550	0.598	0.717	0.657
	BERT-RSP	**0.671**	**0.594**	**0.579**	**0.630**	0.703	0.649
	BERT-MCA	0.588	0.530	0.512	0.557	**0.746**	**0.682**
QCs	BERT	0.340	0.590	0.390	0.431	0.570	0.519
	BERT-RSP	0.334	0.687	0.420	0.450	0.526	0.493
	BERT-MCA	**0.428**	**0.702**	**0.503**	**0.532**	**0.596**	**0.546**
QC	BERT	0.556	0.669	0.590	0.607	0.778	0.752
	BERT-RSP	0.613	0.702	0.641	0.655	0.733	0.720
	BERT-MCA	**0.754**	**0.821**	**0.758**	**0.786**	**0.867**	**0.811**

First, all models achieve high scores across all metrics over QA and QpA queries, indicating their ability to accurately retrieve relevant information; clearly, this is not surprising since QA and QpA queries contain information seen during the models' training. More

Table 4.4 Performance results of task-adaptive pre-trained LegalBERT models and comparison with base LegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	*Rec*	*Prec*	F^{μ}	F^{M}	*Rec*@3	*MRR*
QA	LegalBERT	0.996	0.987	0.990	0.991	0.997	0.993
	LegalBERT-RSP	**0.998**	**0.995**	**0.996**	**0.996**	0.995	0.995
	LegalBERT-MCA	**0.998**	0.991	0.994	0.995	**0.998**	**0.996**
QpA	LegalBERT	0.975	0.944	0.951	0.959	0.971	0.962
	LegalBERT-RSP	0.957	0.924	0.929	0.941	0.961	0.949
	LegalBERT-MCA	**0.980**	**0.952**	**0.959**	**0.966**	**0.982**	**0.969**
QR	LegalBERT	**0.703**	**0.674**	**0.658**	**0.688**	**0.804**	**0.740**
	LegalBERT-RSP	0.665	0.627	0.603	0.646	0.746	0.686
	LegalBERT-MCA	0.663	0.625	0.597	0.643	0.768	0.697
QpR	LegalBERT	**0.707**	**0.685**	**0.662**	**0.696**	**0.775**	**0.726**
	LegalBERT-RSP	0.697	0.675	0.645	0.686	0.754	0.694
	LegalBERT-MCA	0.654	0.613	0.586	0.633	0.754	0.692
QCs	LegalBERT	0.357	0.668	0.419	0.465	0.548	0.482
	LegalBERT-RSP	0.312	0.678	0.402	0.427	0.489	0.455
	LegalBERT-MCA	**0.426**	**0.692**	**0.493**	**0.527**	**0.629**	**0.560**
QC	LegalBERT	**0.794**	**0.862**	**0.806**	**0.826**	0.889	0.848
	LegalBERT-RSP	0.624	0.702	0.641	0.661	0.822	0.732
	LegalBERT-MCA	0.754	0.810	0.764	0.781	**0.911**	**0.860**

Table 4.5 Performance results of task-adaptive pre-trained EULegalBERT models and comparison with base EULegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	*Rec*	*Prec*	F^{μ}	F^{M}	*Rec*@3	*MRR*
QA	EULegalBERT	0.984	0.934	0.946	0.958	0.965	0.962
	EULegalBERT-RSP	**0.996**	**0.990**	**0.992**	**0.993**	**0.998**	**0.995**
	EULegalBERT-MCA	0.986	0.965	0.973	0.975	0.986	0.983
QpA	EULegalBERT	0.878	0.793	0.812	0.833	0.900	0.870
	EULegalBERT-RSP	**0.955**	**0.942**	**0.939**	**0.948**	**0.970**	**0.957**
	EULegalBERT-MCA	0.921	0.877	0.881	0.898	0.955	0.927
QR	EULegalBERT	0.519	0.483	0.451	0.501	0.609	0.557
	EULegalBERT-RSP	**0.689**	**0.667**	**0.633**	**0.678**	**0.754**	**0.691**
	EULegalBERT-MCA	0.665	0.624	0.605	0.644	0.674	0.651
QpR	EULegalBERT	0.371	0.325	0.291	0.347	0.558	0.478
	EULegalBERT-RSP	**0.648**	**0.608**	**0.585**	**0.627**	**0.754**	**0.682**
	EULegalBERT-MCA	0.638	0.571	0.566	0.603	0.681	0.657
QCs	EULegalBERT	0.215	0.425	0.262	0.285	0.349	0.327
	EULegalBERT-RSP	0.344	0.670	0.425	0.455	0.533	0.491
	EULegalBERT-MCA	**0.391**	**0.730**	**0.453**	**0.509**	**0.544**	**0.506**
QC	EULegalBERT	0.385	0.474	0.399	0.425	0.689	0.558
	EULegalBERT-RSP	0.603	0.724	0.644	0.658	**0.778**	0.750
	EULegalBERT-MCA	**0.763**	**0.882**	**0.801**	**0.818**	**0.778**	**0.781**

challenging are the QR/QpR and QCs/QC queries which are based on contents from recitals and commentaries, respectively. Generally, LegalBERT consistently outperforms BERT,

which would indicate the benefits of adapting the models to the legal domain prior to the GDPR article retrieval task. However, EULegalBERT shows significantly lower performance compared to the other two models. This performance gap should be ascribed to the fact that EULegalBERT results from a further pre-training of BERT over EU specific resources, which limited knowledge expansion w.r.t. that gained by LegalBERT over a much larger set of legal corpora in relation to BERT.

Impact of task-adaptive pre-training. Let us now focus on the impact of task-adaptive pre-training on the GDPR article retrieval task, whose results are summarized in Tables 4.3–4.5; on each of those tables, we also report the scores achieved by the corresponding model without task-adaptive pre-training (cf. Table 4.2).

Considering first BERT models (Table 4.3), we find that BERT-MCA achieves the highest scores for most query types, and the advantage over the other two models are particularly evident for the most difficult query sets. BERT-RSP generally performs better than the base BERT in terms of precision, recall, and f-measures, with the exception of QA/QpA query sets, which would indicate that task-adaptation based on the RSP pre-training task can even worsen performance on article retrieval when the query content does not deviate much from the training data. Moreover, BERT-RSP consistently falls behind BERT-MCA, which highlights the superiority of the latter form of task-adaptive pre-training when applied to BERT.

As reported in Table 4.4, task-adaptive pre-training of LegalBERT showcases quite different trends from the BERT counterpart. While MCA maintains a significant advantage over RSP especially for the commentary-based queries (i.e., QCs and QC), the same does not hold for the other queries, especially the recital-based queries (i.e., QR and QpR). Also, on the latter query types and on QC, both variants of task-adaptive pre-training are not able to improve performance over the base LegalBERT. This would suggest MCA (and RSP) might not necessarily take advantage when, as it is the case for LegalBERT, the base model has a pre-training knowledge on a legal domain that would enclose the targeted one (i.e., GDPR in our setting).

Table 4.5 shows how EULegalBERT appears to take advantage when task-adapted via RSP, on query sets QA- QpR, or via MCA, on commentary-based queries. However, compared to the previous results, EULegalBERT models are still outperformed by LegalBERT and BERT models (apart from very few exceptions, such as precision on QCs and QC queries).

Overall, our findings suggest that task-adaptive pre-training, especially based on MCA, can yield better results when applied to base BERT and its legal pre-trained models, and this particularly holds in terms of *Rec*@3 and *MRR* criteria.

4.3.2 Training with recital data enrichment

Table 4.6 shows results corresponding to the recital-enriched models. For each query set and criterion, the table reports the percentage variation (i.e., increase/decrease) achieved by a recital-enriched model w.r.t. its corresponding non-recital-enriched model; moreover, the last row of each query-set subtable shows the absolute best-performing model, considering both the recital-enriched models and the previously analyzed models.

First, we notice that while the improvements are negligible for the QA/QpA query types, some significant increase in performance holds for the QCs/QC query types, particularly for the BERT and EULegalBERT models. Also, it clearly does not come to our surprise that leveraging recitals is beneficial for all domain-adaptive and task-adaptive pre-trained models when evaluated on recital-based queries (i.e., QR and QpR). In such cases, the absolute best model is LegalBERT-r, which also indicates that task-adaptive pre-training is not needed for the target task as its lack can be well compensated by the recital data enrichment.

More importantly, task-adaptive pre-training, especially based on MCA, reveals to be essential to maximize performance in relation to the non-recital-based query sets. Moreover, a combination of MCA and recital-enrichment leads to the absolute best model for the QCs type. This is another remarkable aspect since it supports our intuition of the beneficial effect of integrating the complementary role of recitals for the GDPR articles into task-adaptive pre-training of models to be fine-tuned for the article retrieval task.

Table 4.6 Percentage variations of the recital-enriched BERT-r, LegalBERT-r, and EULegalBERT-r w.r.t. their respective base models (i.e., BERT, LegalBERT, and EULegalBERT). The absolute best-performing model is reported in the last row of each query-set subtable.

Query type	Model	*Rec*	*Prec*	F^{μ}	F^{M}	*Rec*@3	*MRR*
QA	BERT-r	-0.51%	+0.28%	-0.28%	-0.11%	-0.15%	-0.49%
	BERT-RSP-r	+0.39%	+0.79%	+0.69%	+0.59%	+0.46%	+0.69%
	BERT-MCA-r	+0.27%	+0.13%	+0.28%	+0.20%	+0.31%	+0.31%
	LegalBERT-r	-0.67%	+0.03%	-0.58%	-0.32%	0%	-0.39%
	LegalBERT-RSP-r	-0.34%	-0.16%	-0.28%	-0.25%	-0.15%	-0.22%
	LegalBERT-MCA-r	-0.27%	-0.14%	-0.20%	-0.21%	-0.30%	-0.26%
	EULegalBERT-r	+0.14%	+3.47%	+2.69%	+1.82%	+2.67%	+1.88%
	EULegalBERT-RSP-r	+1.32%	+6.17%	+5.04%	+3.75%	+3.14%	+3.27%
	EULegalBERT-MCA-r	+1.13%	+5.84%	+4.72%	+3.49%	+3.46%	+3.20%
	LegalBERT-RSP (-MCA)*	0.998	0.995	0.996	0.996	0.998*	0.9965*
QpA	BERT-r	-4.03%	-6.46%	-5.91%	-5.27%	-0.16%	-1.42%
	BERT-RSP-r	-0.48%	-1.27%	-0.72%	-0.88%	+0.64%	+1.20%
	BERT-MCA-r	-1.19%	-2.96%	-3.08%	-2.09%	+0.16%	-1.14%
	LegalBERT-r	-3.46%	-4.39%	-4.29%	-3.93%	+0.31%	-0.79%
	LegalBERT-RSP-r	-1.30%	+0.10%	-0.96%	-0.59%	+0.16%	-0.68%
	LegalBERT-MCA-r	-1.89%	-1.62%	-1.59%	-1.75%	-0.16%	-0.63%
	EULegalBERT-r	-10.23%	-2.18%	-7.21%	-6.17%	-2.86%	-3.78%
	EULegalBERT-RSP-r	+8.46%	+16.85%	+14.61%	+12.71%	+7.23%	+9.25%
	EULegalBERT-MCA-r	+8.96%	+16.98%	+14.79%	+13.03%	+7.40%	+8.66%
	LegalBERT-MCA	0.980	0.952	0.959	0.966	0.982	0.969
QR	BERT-r	+41.83%	+51.99%	+57.80%	+46.92%	+34.31%	43.83%
	BERT-RSP-r	+5.89%	+7.92%	+8.57%	+6.94%	+5.88%	+4.08%
	BERT-MCA-r	+4.64%	+2.39%	+4.30%	+3.46%	+10.78%	+5.22%
	LegalBERT-r	+34.55%	+39.80%	+43.14%	+37.18%	+23.42%	+32.69%
	LegalBERT-RSP-r	+4.10%	+4.46%	+3.95%	+4.28%	-2.70%	-0.19%
	LegalBERT-MCA-r	+12.09%	+12.51%	+12.29%	+12.31%	+2.70%	+5.13%
	EULegalBERT-r	+82.15%	+95.01%	+108.48%	+88.59%	+63.10%	+76.29%
	EULegalBERT-RSP-r	+29.58%	+31.65%	+35.17%	+30.64%	+22.62%	+22.64%
	EULegalBERT-MCA-r	+44.10%	+47.45%	+51.29%	+45.81%	+26.19%	+29.00%
	LegalBERT-r	**0.946**	**0.942**	**0.941**	**0.944**	**0.993**	**0.982**
QpR	BERT-r	+48.05%	+60.76%	+66.33%	+54.41%	+36.36%	+46.15%
	BERT-RSP-r	+11.29%	+9.67%	+11.90%	+10.44%	+6.06%	+7.05%
	BERT-MCA-r	+8.93%	+12.44%	+11.81%	+10.73%	+11.11%	+7.53%
	LegalBERT-r	+35.72%	+38.70%	+43.24%	+37.22%	+28.04%	+35.30%
	LegalBERT-RSP-r	+4.08%	+4.33%	+4.48%	+4.21%	+1.87%	+2.88%
	LegalBERT-MCA-r	+1.39%	-1.83%	-2.12%	-0.27%	+6.54%	+4.51%
	EULegalBERT-r	+150.46%	+180.98%	+212.85%	+165.86%	+74.03%	+100.73%
	EULegalBERT-RSP-r	+80.38%	+93.20%	+107.180	+87.00%	+31.17%	+42.34%
	EULegalBERT-MCA-r	+95.110%	+110.51%	+129.21%	+103.03%	+32.47%	+48.59%
	LegalBERT-r	**0.960**	**0.950**	**0.949**	**0.955**	**0.993**	**0.982**
QCs	BERT-r	-1.20%	+9.66%	+1.84%	+2.51%	-6.45%	-9.89%
	BERT-RSP-r	-0.21%	+14.13%	+9.03%	+4.59%	-1.94%	-0.21%
	BERT-MCA-r	+17.88%	+16.34%	+23.19%	+17.31%	+8.39%	+4.63%
	LegalBERT-r	+11.43%	+6.32%	+10.03%	+9.60%	0%	+2.87%
	LegalBERT-RSP-r	-0.70%	-2.51%	+5.24%	-1.34%	+10.74%	+8.78%
	LegalBERT-MCA-r	+26.91%	+15.20%	+25.25%	+22.57%	+12.08%	+17.86%
	EULegalBERT-r	-30.57%	-7.88%	-28.59%	-24.314	-16.842	-18.575
	EULegalBERT-RSP-r	+105.58%	+80.23%	+97.89%	+96.32%	+64.21%	+69.46%
	EULegalBERT-MCA-r	+67.82%	+60.49%	+70.08%	+65.29%	+65.26%	+61.23%
	LegalBERT-MCA-r	**0.452**	**0.770**	**0.524**	**0.570**	**0.614**	**0.569**
QC	BERT-r	+13.88%	+5.64%	+10.04%	+9.99%	+8.57%	-2.21%
	BERT-RSP-r	+13.06%	+19.29%	+15.70%	+15.80%	+2.86%	-1.00%
	BERT-MCA-r	+15.71%	+4.06%	+11.09%	+10.12%	+8.57%	+5.19%
	LegalBERT-r	-6.57%	-6.54%	-7.28%	-6.56%	+2.50%	-2.56%
	LegalBERT-RSP-r	-2.71%	-3.22%	-2.82%	-2.96%	+2.50%	-1.29%
	LegalBERT-MCA-r	-6.86%	-5.16%	-5.09%	-6.05%	+2.50%	-2.38%
	EULegalBERT-r	-7.94%	+11.56%	-0.61%	-0.11%	-22.58%	-15.89%
	EULegalBERT-RSP-r	+101.77%	+75.89%	+94.54%	+89.27%	+22.58%	+45.16%
	EULegalBERT-MCA-r	+61.18%	+54.10%	+64.33%	+57.93%	+12.90%	+36.89%
	LegalBERT (-MCA)*	0.794	0.862	0.806	0.826	0.911*	0.860*

Chapter 5

Network Analysis of Law Article References

Research in artificial intelligence and law has traditional focused on a number of problems that are typically addressed by natural language processing and machine learning models and methods. Despite a relative gap with respect to the network science discipline, a few works have been proposed to investigate the complexity in a legal corpus domain by leveraging network analysis methods. For instance, Fowler et al. [32] use a network-based representation to characterize the most important precedents at the U.S. Supreme Court. Moser et al. [68] study the structural properties of the citation networks inferred from Austrian Supreme Court decisions. Koniaris et al. [46] model the legislation network of the European Union law, and explore its topological structure and evolution over time, also evaluating its resilience. Mazzega et al. [64] study the network of French Legal codes. Moreover, Zhang et al. [113] propose a software tool to visually explore semantic-based citation networks to ease the analysis of the relations and the evolution of legal issues. To the best of our knowledge, the Italian Civil Code has not been studied through its article-reference-based structure so far. Therefore, our main goal in this work is to contribute with a study of the networks that can be inferred from the ICC corpus based on article references. This allows us to extend our scope beyond the canonical exploration boundaries of the legal domain, by providing new insights on the interpretation and evaluation of the Italian civil law. In this chapter, we explore the Italian Civil Code (ICC) from an unprecedented perspective based on network analysis. We develop a text processing method to identify and extract article references from the ICC and, upon these, we define network models capturing their relation structures either at a book and corpus scale. The exploitation of the main structural features of these networks leads us to unveil meaningful patterns, holding within and across the books composing the ICC. Furthermore, by leveraging a community detection task, we investigate whether the

formation of a community is related to the topic coherence of its assigned articles over the portions of books involved. Our findings reveal useful indicators that may help legal experts and practitioners enhance their knowledge from a novel perspective provided by the network of article references through the ICC.

5.1 Data modeling

The ICC article references can naturally be modeled as a network, so to enable the discovery and analysis of relation patterns among the article contents beyond the original, logical organization of the corpus.

5.1.1 Extraction of article references

An article in a specific book of the ICC may contain citations of one or more articles, which are from the same or a different book. We will refer to these citations as *article references*.

Unfortunately, identifying and extracting article references from ICC articles is not straightforward, because of two main reasons:

1. The ICC and its books are not designed as hypertexts, neither any index structure containing article references is originally available.

2. An article may also contain references to laws or legal items that do not correspond to ICC articles.

Therefore, since our goal is to infer citation networks from the lists of article references that are contained in each ICC book, we developed an approach to identify and extract *valid* article references, i.e., citations of articles within the ICC only. In the following, we elaborate on the text pre-processing steps that were carried out for the task at hand.

The ICC is obviously publicly available, in various digital formats. From one of such sources, we extracted the contents of each article from all books. Note that we normalized all variants and abbreviations of frequent keywords such as “articolo” (i.e., article), “decreto legislativo” (i.e., legislative decree), “Gazzetta Ufficiale” (i.e., Official Gazette), and finally we lowercased all letters.

An article reference is comprised of three parts:

- *Prefix*, which corresponds to a common root for all lexical variants of article; typically, the abbreviation “*art*” is used.

- *Article id*, which follows the numerical intervals specific to each book (as described earlier in this section).
- *Variant suffix*, which is optional and used to designate a variant of a given article, which is however not alternative, and hence must be treated as a separate article in the book; specifically, an article variant suffix corresponds to a Latin adverbial numeral, to express the multiplicity of occurrence, and hence subsequent versions of a given article id, i.e., "*bis*" (stands for "twice"), "*ter*" ("three times"), "*quater*" ("four times"), and so on.

As mentioned before, in addition to references to other articles in the ICC, an article may contain references to legal items that are external to the ICC corpus. Since they appear in similar textual format, it is not trivial to distinguish between the two types of references. Nonetheless, in the article contents, we can recognize *cue-words* to exploit as hints for the presence or not of references of either type:

- *Relevant* cue-words are used to characterize the presence of valid article references; for instance, "*codice civile*" ("civil code") and its lexical variants.
- *Irrelevant* cue-words are instead used to locate references to external sources. These include "*legge*" ("law"), "*decreto legislativo*" ("legislative decree"), "*sentenza*" ("judgment"), and are usually followed by a date in some format. Moreover, they also include other words, such as "*comma*" ("paragraph"), which may follow a reference inside a portion of the context delimited by round brackets.

It should be noted that relevant cue-words are much fewer and less frequently used than irrelevant cue-words. Moreover, there are cases in which both types of cue-words are found within an article content, also with multiple occurrences of one or both types, while in other cases they are both missing. For example, the following is an excerpt from article 42-bis that contain multiple valid references:

```
"[...] si applicano inoltre art2499, art2500, art2500-bis, art2500-ter, secondo
comma art2500-quinquies, art2500-nonies, in quanto compatibili. [...]"
```

By contrast, in the next example from article 86, references are to external sources only, and hence must be discarded from our analysis:

```
"[...] la legge 20 maggio 2016, 76 ha disposto (con art1, comma 35) che la presente
modifica acquista efficacia dal 5 giugno 2016. [...]"
```

Algorithm 1: Extraction of article references

Data: Set of articles $\mathscr{A}$ in the ICC; list of relevant cue-words *art-refs_words* and list of irrelevant cue-words *ext-ref_words*
Output: Article reference sets AR

```
AR ← {}
foreach a ∈ 𝒜 do
    AR_a ← {}
    contexts ← getContexts(a)
    foreach c ∈ contexts do
        isIrrelevant ← search(ext-refs_words, c) and not search(art-refs_words, c)
        if not isIrrelevant then
            ⟨a, refs⟩ ← findAll("art[0-9]+[a-z]*", c)
            AR_a ← AR_a ∪ {⟨a, refs⟩}
        end
    end
    AR ← AR ∪ AR_a
end
```

To process the article contents for the task of article reference extraction, we first segmented an article's text into shorter passages, dubbed contexts, that are delimited by "." or ";" and preceded by an alphanumerical sequence of at least four characters; the latter constraint is needed to avoid selecting false contexts, e.g., corresponding to one of the many abbreviations that are found in the ICC articles such as "d.p.r.", "c.p.c.", "c.p.p.", "etc.".

Algorithm 1 sketches our designed procedure to identify and extract article references from the text of ICC articles. The procedure takes in input two lists of words, which are used to find indicators of presence of relevant and irrelevant *cue-words* for the task at hand.

In the extraction procedure, the *search* function takes in input a list of cue-words and a context to check whether any cue-word occurs within the context. If irrelevant cue-words are found in the context but relevant cue-words are not, then the next context is processed; otherwise, the context is scanned left-to-right through the *findAll* function, which returns all non-overlapping matches of the article-references pattern in the given context.

By applying Algorithm 1 to the aforementioned examples, the output for the first text will be a list of 6 references to associate with article 42-bis, while in second text the output is an empty list since the occurrence of word "art1" is correctly recognized as an external, hence irrelevant, reference, due to the presence of irrelevant cue-words (i.e., "la legge 20 maggio 2016" and "comma 35").

Table 5.1 Statistics about the extraction of the article references from the ICC books.

	Book-1	Book-2	Book-3	Book-4	Book-5	Book-6
# articles	395	345	364	891	713	331
# articles w/ references	123	71	45	77	258	88
# article-references	243	132	70	118	551	180
avg. # article-references per-article	0.615	0.383	0.192	0.132	0.773	0.544
avg. # article-references per-article w/ references	1.976	1.859	1.556	1.532	2.136	2.045

5.1.2 Network models

Let us denote with $\mathscr{A}$ the set of articles in the ICC, and with $\mathscr{B}$ a partition of $\mathscr{A}$ into six groups, each corresponding to a book in the ICC, i.e., $\mathscr{B} = \bigcup_{i=1..6} \mathscr{A}_i$, such that $\mathscr{A}_i \cap \mathscr{A}_j = \emptyset$, for all $i, j \in \{1, \ldots, 6\}$ with $i \neq j$, where $\mathscr{A}_i \subset \mathscr{A}$ is the subset of articles assigned to book i. Let also $r : \mathscr{A} \mapsto 2^{\mathscr{A}}$ denote a function that associates each article a to a set of articles that are referred to by a, possibly including articles that are from the same book of a or a different one.

Based on the article-reference relation, and by differentiating at level of single book or entire corpus, we define the following network models of interest to the study of the ICC.

Book-induced networks. Our first defined network model focuses on representing the relations between articles of each particular book and their article references, which may cross the boundary of the book itself. Given a book, the induced network is a directed graph built from the article-reference list of all articles of that book. Formally, for each book i, we define the book-induced network for i as the directed graph $G_i = \langle V_i, E_i \rangle$ such that $V_i = \{a \in \mathscr{A}_i \mid r(a) \neq \emptyset\} \cup \{a \in \mathscr{A} \setminus \mathscr{A}_i \mid \exists a' \in \mathscr{A}_i, a \in r(a')\}$ and $E_i = \{(a, a') \mid a, a' \in V_i, a \in r(a')\}$.

Global or **corpus network.** Our second network model encompasses all relations among articles observed in the entire ICC. Therefore, we define the ICC corpus network $G_{\mathscr{B}}$ as the directed graph obtained by merging all book-induced networks, i.e., $G_{\mathscr{B}} = \langle V, E \rangle$ such that $V = \bigcup_{i=1..6} V_i$ and $E = \bigcup_{i=1..6} E_i$.

5.2 Structural Analysis of the ICC Networks

In this section we present our analysis of the book-induced and corpus networks, which is aimed to characterize main structural features of such networks and to unveil interesting patterns underlying the article references within and across books of the ICC. We organize our presentation into two subsections: the first one (Section 5.2.1) is concerned with *macroscopic*

Table 5.2 Summary of structural characteristics of the ICC corpus network (first column) and book-induced networks (subsequent columns).

	$G_{\mathscr{B}}$	G_1	G_2	G_3	G_4	G_5	G_6
#nodes	1 147	223	144	95	161	432	171
#edges	1 294	243	132	70	118	551	180
reciprocity	3.4%	4.9%	3%	2.9%	0%	2.2%	7.8%
density	0.001	0.005	0.006	0.008	0.005	0.003	0.006
average degree*	2.218	2.126	1.806	1.453	1.466	2.523	2.023
average in-degree	1.128	1.090	0.917	0.737	0.733	1.275	1.053
% sources	37.2	35	35.4	41.1	43.5	35.4	31
% sinks	42.3	44.8	50.7	52.6	52.2	40.3	48.5
assortativity*	0.016	-0.184	-0.063	-0.058	-0.141	0.012	-0.173
assortativity	-0.035	-0.198	-0.051	-0.072	-0.158	-0.037	-0.196
average path length	2.241	1.639	1.393	1.104	1.064	2.384	1.568
diameter	7	6	3	3	3	7	5
transitivity*	0.099	0.119	0.135	0.160	0.107	0.098	0.109
clustering coefficient*	0.166	0.227	0.225	0.197	0.220	0.129	0.190
clustering coefficient *(full averaging)**	0.081	0.106	0.097	0.039	0.055	0.072	0.091
#strongly connected components	1 128	218	142	93	161	427	166
#weakly connected components *	157	23	29	30	50	47	23
modularity*	0.891	0.866	0.882	0.914	0.946	0.807	0.876
#communities*	174	32	32	30	50	59	30
modularity	0.892	0.868	0.881	0.909	0.946	0.812	0.876
#communities	175	32	32	30	50	60	30

* Statistic calculated by discarding edge orientation

structural properties of the networks, whereas the second (Section 5.2.2) is focused on *mesoscopic* structural properties based on the outcome of a community detection task.

5.2.1 Macroscopic properties

Table 5.2 reports main results of our macroscopic structural analysis if the ICC corpus and book-induced networks. As a first general remark, only a portion of the articles is actually involved in article reference relations; in particular, considering the global network, the number of nodes (1 147) corresponds to about 36% of the articles within the ICC. This is actually not surprising, since it is commonly expected that a law article should be self-contained or self-explanatory. Nonetheless, there are norms regulating specific sections of the law code which, to be completely specified, require one or more references to different

articles, possibly crossing the boundaries of a book. Indeed, this is what happens for a fraction of the ICC, which is not negligible, and hence deserves to be investigated.

In light of the above remark, one trait common to all the networks is the low *density*, which is the actual number of edges divided by the maximum possible number of links in the network, i.e., $|E|/(|V|(|V|-1))$ for any directed graph $G = \langle V, E \rangle$. Partly related is also the low average *degree*, i.e., the average number of references involving an article, as well as low average *in-degree*, i.e., the average number of references to an article. Focusing on the latter, we observe that the ICC corpus network $G_{\mathscr{B}}$, and the book-induced networks for Book-1, Book-5 and Book-6 have average in-degree above 1 (i.e., on average, each article is pointed by at least one other article), whereas the remaining networks exhibit a value lower than 1, indicating broader isolation. Indeed, by looking at an article's in-degree as a measure of its authoritativeness and usefulness to clarify and deepen the semantics of the articles that point to it, we find out that some articles indeed take a central role in the ICC. Figure 5.1 displays a visualization of the ICC corpus network, where articles of the same book are colored the same, and nodes with highest in-degree are made evident with associated label.

By contrast, the periphery of each of the networks represents a significant fraction of nodes, as it can be noted from the percentages of *source* and *sink* nodes reported in Table 5.2. We refer to source and sink nodes as those having only outgoing links and incoming links, respectively, i.e., source articles include others in their definition but are not referred to by other articles, whereas sink articles are used in the definition of one or more articles but they do not need others to complete themselves.

The above duality also impacts on the degree correlation, also known as *assortativity*, which captures how the probability of links between two nodes is influenced by their degree [71, 72]. For all networks, we observe negative correlation, which means that articles with different degrees tend to link each other.

We also analyzed the tendency of articles to form strong ties with closure patterns, specifically dyadic closure or *reciprocity* and *triadic closure*. The former is computed as the fraction of reciprocal edges, and helps us understanding how articles might reinforce each other to enhance and refine their meaning. Reciprocity is found to be low for all networks, with 3.4% for the corpus network and upper bounded by 8% for the book-induced networks. For the latter, it is interesting to observe two contrasting cases: the one corresponding to zero reciprocity, which characterizes the article references found in Book-4, and the other one corresponding to the maximum reciprocity found in Book-6, thus suggesting a more evident dyadic closure for the articles in that book than in others. As concerns the triadic closure, i.e., the probability of closing connected triplets of nodes, we resorted to both global and local measures [11]. In the former case, we evaluated the the probability that two incident edges

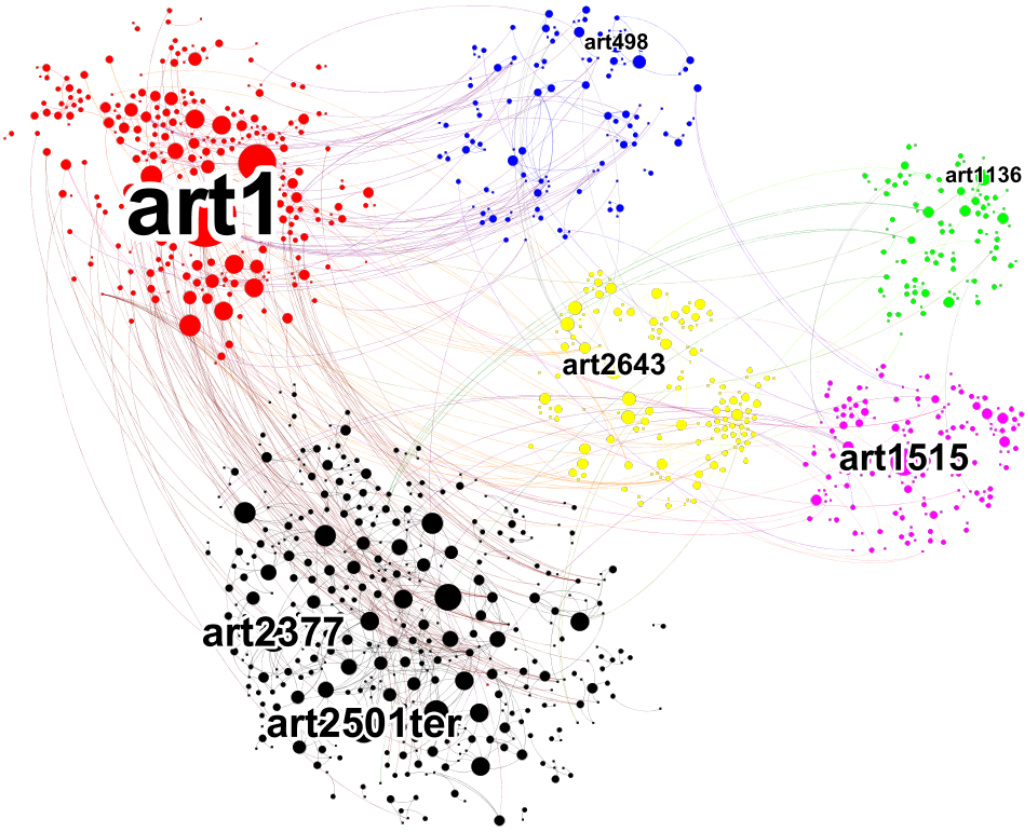


Fig. 5.1 Articles in the ICC corpus network. Node sizes are proportional to the in-degree, and for each book a label representing the article id is associated to the node(s) with highest in-degree. Colors are used to distinguish the six books of the ICC as follows: red (Book-1), blue (Book-2), green (Book-3), magenta (Book-4), black (Book-5), yellow (Book-6).

are completed by a third one to form a triangle, namely *transitivity*. In the latter case, instead, we evaluated the transitivity at node level, i.e., how strongly connected are the neighbors of a node, namely *local clustering coefficient*. Although both measures show low values, it should be noted how the local clustering coefficient shows slightly greater values when ignoring source and sink articles.

We also investigated reachability aspects in the various networks. First, we examined the *average path length*, i.e., the average of the pairwise distances between nodes in a network,

where the distance between two nodes corresponds to the length of a shortest path connecting them. As it can be noted from the table, the average path length is slightly above 2 for the corpus network and for the Book-5 network only. Also, the *diameter*, i.e., the shortest-path distance between the two most distant nodes in the network, is maximum in the Book-5 network (7) and determines the diameter for the corpus network as well; Book-1 and Book-6 networks have diameter equal to 6 and 5, respectively, while the remaining networks have diameter 3.

5.2.2 Mesoscopic properties

Reachability-based approaches to the identification of subnetworks with particular properties of connectivity focus on the existence of paths, regardless of the distance. In this respect, for each of the networks under study, we calculated the *strongly* and *weakly* connected components, which correspond to the maximal subgraphs where every node is reachable from every other node through a directed or undirected path, respectively. Given the outcome of the above discussed analysis steps, it does not come to our surprise that the observed number of strongly connected components is extremely high, which indicates the formation of small groups of articles that are involved in chains of references when accounting for edge orientation, and highlights the lack of *multi-hop references* (i.e., *art. x* refers to *art. y*, which in turn refers to *art. z*). By discarding edge orientation, instead, mutual connectivity clearly increases, and the detected number of weakly connected components is one order of magnitude lower than the node-set size. This trait confirms the existence of strategic articles, namely those referred by other ones and that support (undirected) connectivity between articles.

Connected components can be seen as a raw form of *communities*, based on minimal requirements of connectedness. However, according to a density hypothesis, communities should correspond to locally dense neighborhoods of a network. Therefore, we delved into each of the article reference networks by carrying out a mesoscale-level analysis to shed light on the underlying *community structure*, i.e., a division of a network into regions such that the nodes in each region should be highly linked with each other, whereas few links should exist between the regions. Note however that the amount of edges per se is not an optimal proxy to quantify the community structure: a good community structure is not merely one in which there are few edges between communities, rather it is one in which there are fewer than expected edges between communities. This is the key principle underlying the theory of *modularity* to discover a community structure in a network: intuitively, the modularity of a community is the total difference of the fraction of the edges within the community w.r.t. the expected such fraction if the edges were distributed at random, so that the higher this

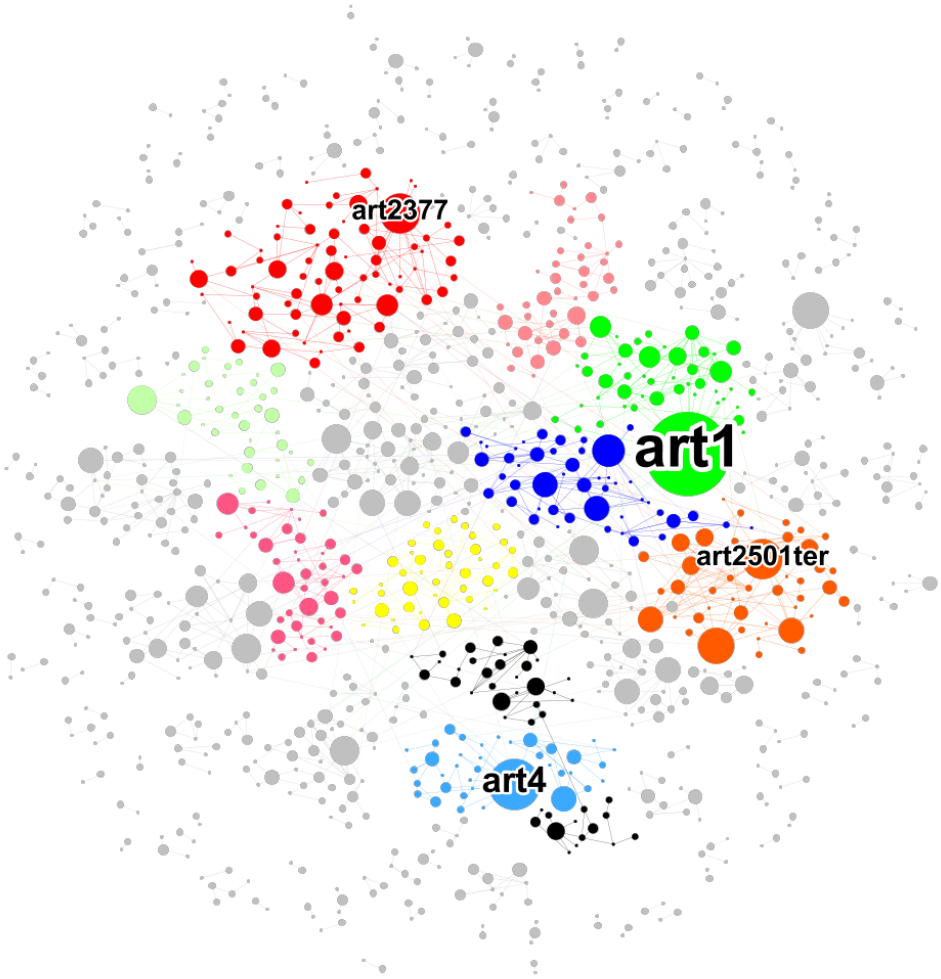


Fig. 5.2 Community structure of the ICC corpus network. Node sizes are proportional to the in-degree, and nodes with in-degree greater than or equal to 10 are labeled with the associated article id. Colors are used to distinguish the top-10 largest communities (nodes assigned to the remaining communities are colored in gray).

deviation, the better the community. In this work, we follow the widely-recognized line of research that resorts to a modularity maximization approach to community detection. In particular, we use the most popular method belonging to this category, namely the *Louvain* method [10], both in its original, undirected version and its variant that accounts for edge orientation while maximizing the modularity.[1]

As reported in Table 5.2, both Louvain and its directed variant lead to the discovery of communities having high modularity (note that modularity is upper bounded by 1) in all

[1] https://github.com/nicolasdugue/DirectedLouvain

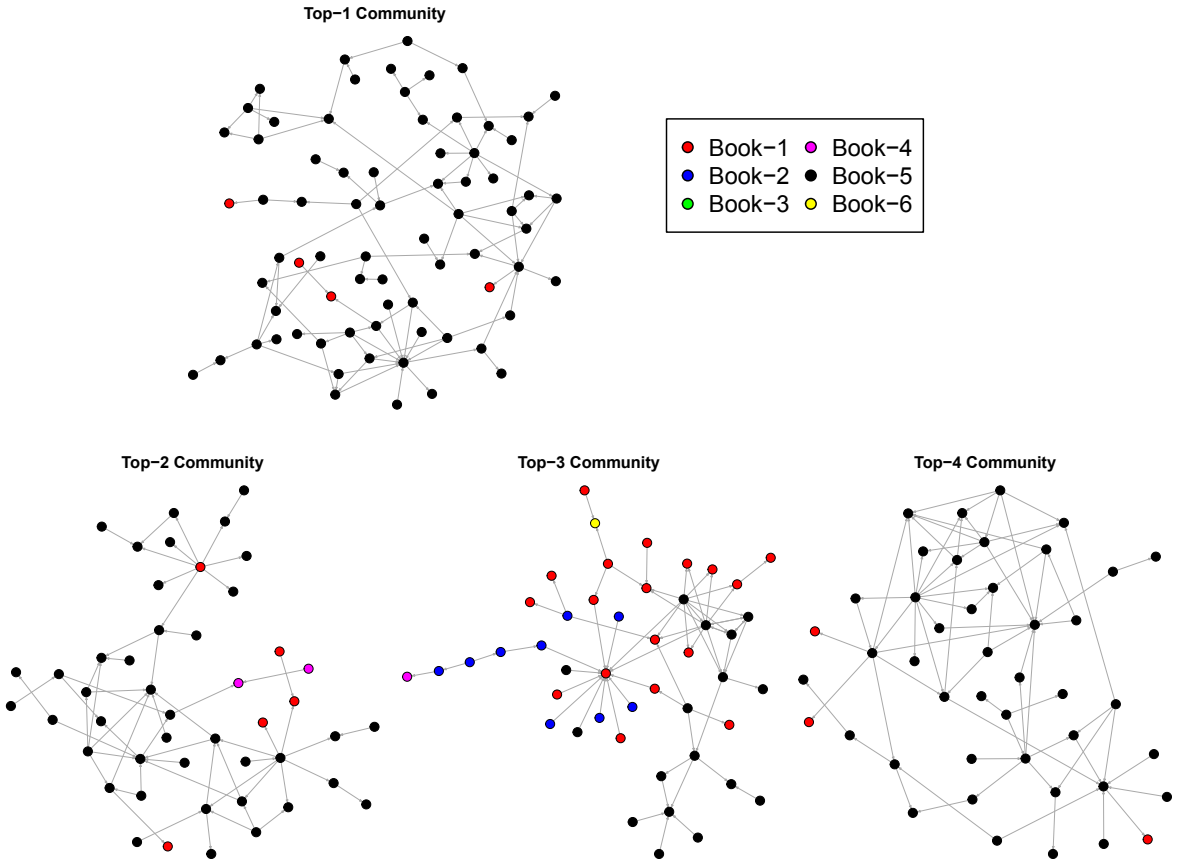


Fig. 5.3 Top-10 largest communities discovered by the Louvain method in the ICC corpus network. Color codes correspond to the ICC books.

networks. This is also consistent with the presence of several connected yet isolated regions within the networks, which are also comparably in size with the connected components previously discussed. Furthermore, it should be noted that accounting for edge orientation does not imply significant differences in modularity as well as number of communities with respect to the outcome of the undirected counterpart. Figure 5.2 illustrates the community structure identified in the ICC corpus network, where the top-10 largest communities are emphasized and distinguished by color. Note that these top-10 communities cover about 40% of the network.

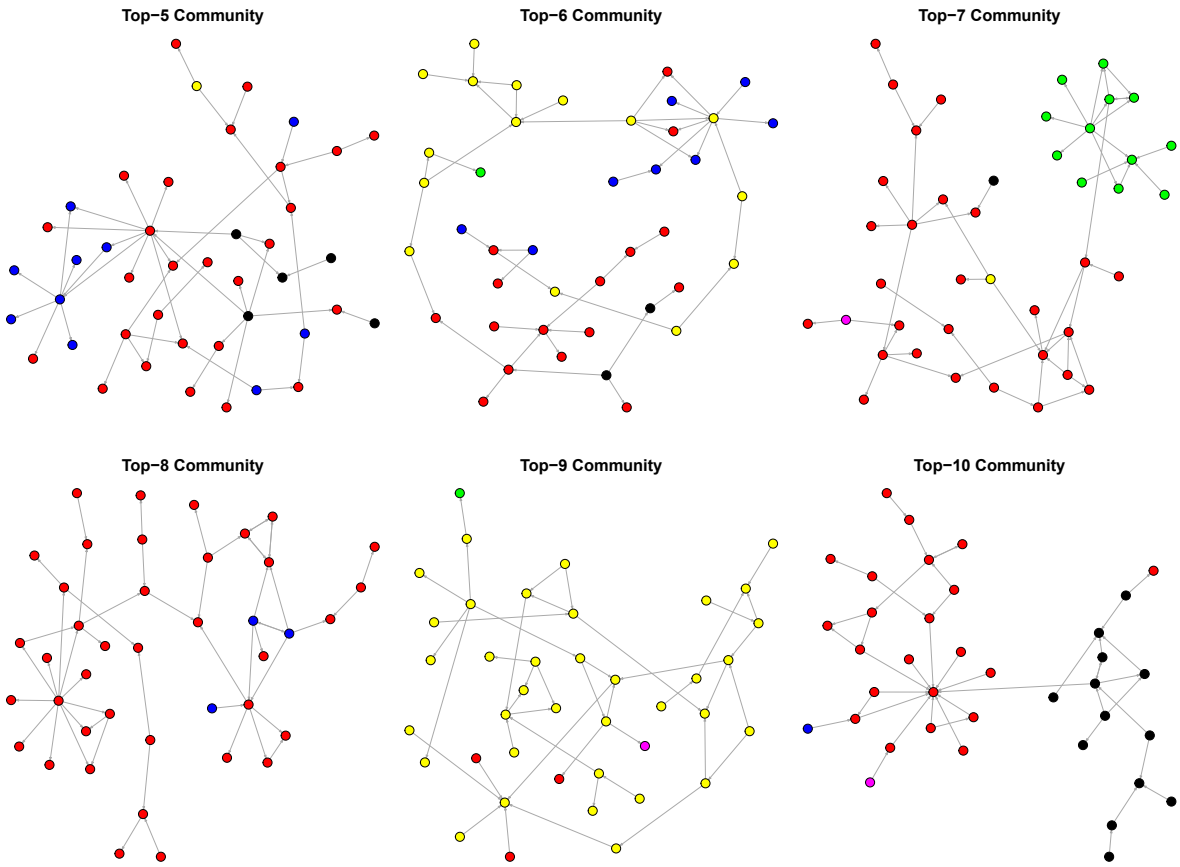


Fig. 5.4 *(Cont.)* Top-10 largest communities discovered by the Louvain method in the ICC corpus network. Color codes correspond to the ICC books.

One interesting question that arises is whether the different communities contain mixed book-memberships, i.e., whether a community may cross the book boundaries through the article reference relations. To this purpose, we explored the top-10 largest communities in the ICC corpus network, which are visualized in Figures 5.3–5.4. As it can be noted from the figures, there are indeed cohesive communities, which are mostly formed by articles from the same book, as well as communities that contain articles from different books. This would suggest that the modular structure in the ICC corpus network can be consistent with either themes of a particular (portion of) book or themes that are shared by (portions of) different books. Clearly, the latter might originate from the opportunity of completing or enhancing the norms provided by a book's article(s) with those provided by other books' article(s).

In this respect, we moved a step forward by exploring the contents of the articles involved in the top-10 communities, in order to gain insights into patterns underlying possible relations

between the formation of communities and the topic coherence. From this analysis stage, several remarks stand out, which we try to summarize as follows:

- Community capturing most representative topics of a particular book. This is likely to happen when the book memberships in the community are mostly or fully homogeneous. For instance, the top-1 community contains articles focused on "administration of the capital of a company", which is a representative topic of Book-5.

- Community unveiling fine-grain topical patterns that are mostly discussed in a book yet complemented with references to other book(s). This refers to sort of strong ties that are formed through article references that cross the boundaries of two or more books. For instance, top-8 and top-10 communities are such a type of community. In the top-8 community, an interesting pattern is found out for "succession" norms (Book-2) in a context of "marital separation" (Book-1).

- Community induced from reinforcement of topic(s) from a book with related topics that differently contribute to the contents of other books. This type of community turns out to be characterized from mixed book-memberships that are distributed over substructures built upon across-book article references. For instance, the top-6 community contains node-articles that belong to five different books; moreover the topic "transcription" that is largely discussed in Book-6 (which is the mostly covered in the top-6 community) is reinforced with the topic "properties", which is shared between Book-1 and Book-2, jointly with the topic "community", which is discussed within Book-1.

In light of the above remarks, we can conclude that a mesoscopic view like that supplied by our discovered community structure can represent a valuable support to enhance the macroscopic (i.e., at book level) or microscopic (i.e., at article level) views that are primarily considered from legal experts and practitioners.

5.3 LawNet-Viz: Visually Exploration of Law Article References Networks

In this section, we present LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. LawNet-Viz is designed to support legal research tasks, hence to help legal professionals searching for referred and semantically related articles, but it also provides a reliable form of legal aid

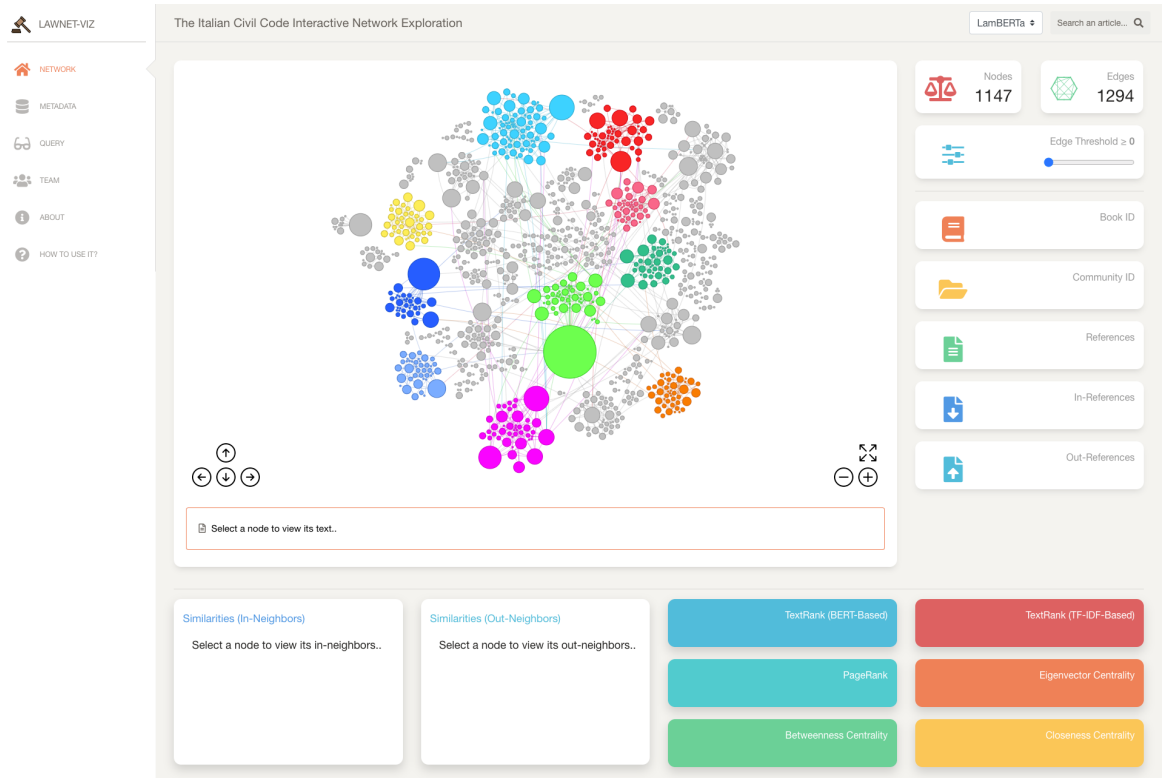


Fig. 5.5 Screenshot of the main interface of LawNet-Viz, with node (i.e., article) colors coding community memberships.

to those who are not familiar with the legal domain. The network of article references can be analyzed at macro-, meso- and microscopic levels, enabling inspection of an article's neighborhood and the associated metadata, along with an extensive set of network statistics including node centrality and ranking measures. Moreover, LawNet-Viz allows the user to refine a network according to the semantic similarity of linked articles, which is captured based on sparse vectorial representations, topic models, word embeddings or deep contextualized embeddings. The usefulness of LawNet-Viz lies in the capability of providing the user with a responsive and interactive visual exploration of law article reference networks, thus reducing the gap her/his knowledge of legal matters — while guaranteeing a fluid and effective user experience. Indeed, LawNet-Viz has the potential of enhancing the ability of lawyers to identify relevant statutes in a cost- and time-effective manner, thereby increasing access to justice. Also, in civil law countries, it can benefit jurists in drafting new codes or abrogating existing ones, by inferring semantic relations between legal documents. For citizens, LawNet-Viz can be helpful to serve their search and consultation needs. Clearly, LawNet-Viz is not intended as a substitute for a full exploration of the social and ethical considerations related to automating legal research, which is beyond the scope of this paper. LawNet-Viz is a system prototype that is designed and implemented to be easily adapted to any law code system. To demonstrate LawNet-Viz, we show its application to the Italian Civil Code (ICC), which exploits the LamBERTa model trained on the ICC in Chapter 3 3.

5.3.1 Design

In this section we provide a conceptual overview of LawNet-Viz and discuss the main functionalities and supported tasks.

LawNet-Viz is logically separated into two independent yet complementary modules that are designed for network and text processing, respectively. The first one is responsible for extracting the references from the articles in the input law corpus to build a law reference network. The resulting graph can be processed through any network analysis software to produce a feature-rich network (i.e., the topology with associated metadata and statistics), which is made available in the form of a JSON object. The second module leverages natural language processing techniques to represent the textual contents of the articles; sparse vectorial representations and deep contextualized embeddings trained on the law corpus are two available but not exclusive options in LawNet-Viz.

The two modules communicate with a third module that is in charge of *(i)* integrating the network topology and metadata with the textual content information, which is used to determine the semantic affinity of any two linked articles, and *(ii)* interactive rendering to enable visual exploration of the integrated data.

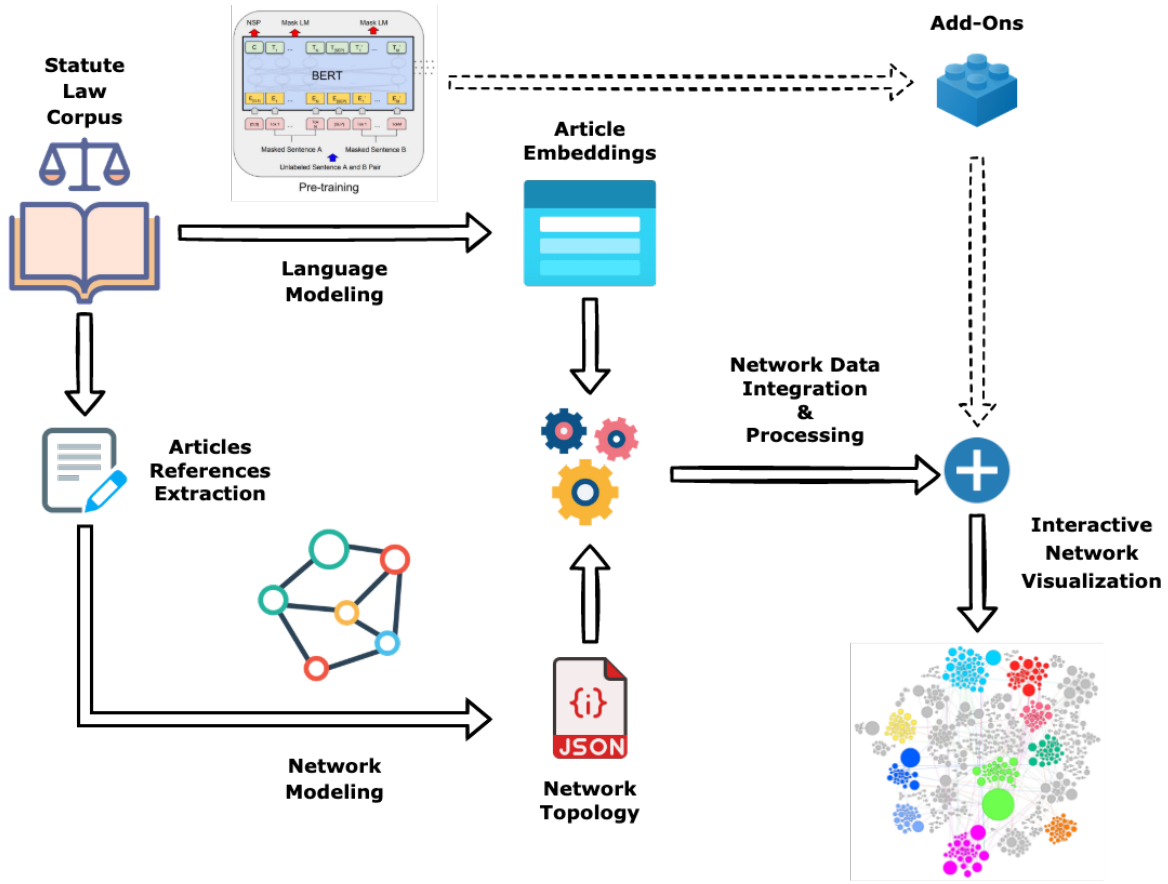


Fig. 5.6 Overview of the architecture of LawNet-Viz.

LawNet-Viz is designed to be a flexible tool not just in terms of the input law code and relating metadata to build a feature-rich article reference network; it can also play a role as front-end for statute retrieval and question-answering modules. In this regard, we allow extending the core functionalities of LawNet-Viz through additional modules, such as the search and retrieval component.

We sketch the workflow of LawNet-Viz in Figure 5.6. Next, we delve into the interaction *modes* of LawNet-Viz to describe the main functionalities available to the user.

Article references extraction Statute law corpora are not commonly available as hypertexts, therefore a text processing method might be required to identify and extract the article references. We address this task in LawNet-Viz and provide an implementation tailored to the specific syntax (i.e., cue-words, abbreviations, etc.) used to denote article references in the input law code. The reference extractor must also be able to recognize and distinguish article references internal to the code from references to external sources; in our implementation, we discard the external references.

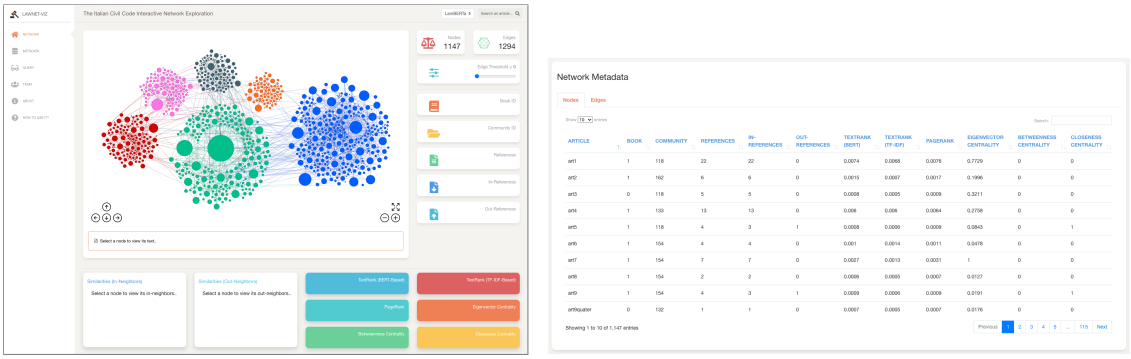


Fig. 5.7 Screenshots of LawNet-Viz network mode, with node (i.e., article) colors coding book memberships (left), and metadata mode (right).

Network mode The main interface of LawNet-Viz corresponds to a view on the article reference network and provides a number of functionalities to explore and manipulate the network (Figures 5.5 and 5.7). By default, the user is provided with a display of the entire network inferred from the analyzed corpus, along with a set of topological statistics. Relations between articles are weighted according to their content similarity; to capture different semantic relation nuances between articles, the user can decide which similarity function to use, and filter edges according to a threshold.

To meet sound user-experience principles, we offer the user a responsive and interactive interface, including but not limited to zooming-in/out and navigating the network through specific controls or using a hand-held pointing device, and selecting articles directly from the network or from a search box. When the user selects an article to delve into details, the article will be highlighted in the network object container along with its 2-hop expanded neighborhood, thus enabling a visual inspection of the article within its surrounding context. Simultaneously, LawNet-Viz will display all available topological and content related information concerning the article, in particular: *(i)* the article's location in the corpus, which allows the user to notice the high-level subdivision of the corpus (e.g., book) a given article belongs to; *(ii)* the count of incoming and outgoing article's references, and a ranking of these references based on the semantic similarity w.r.t. the article according to the selected similarity model function; *(iii)* statistics on node centrality of the article, to gain an insight into the prominence of the article within the network according to different centrality notions, including standard path-based and eigenvector-based methods, as well as content-biased methods.

Metadata mode LawNet-Viz provides a table-based interaction mode to explore the metadata associated with nodes (i.e., articles) and edges (i.e., references), which are organized

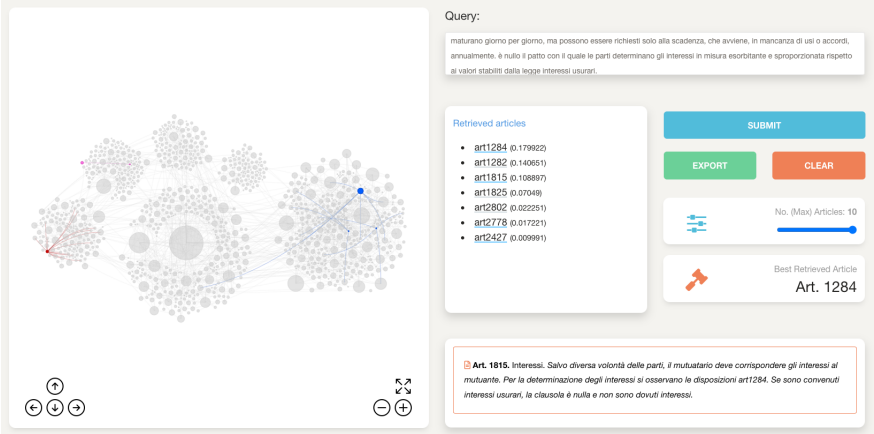


Fig. 5.8 Screenshot of LawNet-Viz query mode.

into two tabs (Figure 5.7). Again, the user experience is highly considered, as we enable responsive table exploration through its search, filter, and sort capabilities, based on the article names and/or associated statistics.

Query mode This is conceived as one add-on of LawNet-Viz, which is designed to fulfill a primary need in exploring a legal corpus, i.e., to retrieve articles relevant to a given input query. Here the user submits a natural language query — which is assumed to be free of references to any article identifier in the law code — and specifies a number k. LawNet-Viz will retrieve the top-k articles from the corpus that satisfy the given query, which are ranked according either to a TF-IDF-like retrieval model or to a relevance probability distribution produced by a contextualized language model specialized for the legal domain. The user can then inspect the retrieved articles, locate their position and context in the reference network, and save the results (Figure 5.8).

5.4 Implementation

In this section we elaborate on our development of the constituting components of LawNet-Viz.

Language modeling To represent the article contents, we cleaned up the text from non-ASCII characters, removed dates and numbers, including those used as article references, and normalized all variants and abbreviations of terms referring to legislation norms or to the code subdivisions (e.g., chapters, paragraphs, etc.).

We consider different types of text representation models: *(i)* sparse vectorial model (based on TF-IDF weighting function), *(ii)* statistical topic model (Latent Dirichlet Allocation), *(iii)* context-free word embeddings (Word2Vec, FastText), and *(iv)* pre-trained contextualized language model based on BERT [25]. For the implementation of the first three types of models we used the free open-source Python library Gensim [118], whereas for the BERT model(s) we resorted to the HuggingFace Transformers library;[2] in this regard, although any out-of-the-box BERT model can be used, we are aware of the existence of recently developed BERT-based models which are adapted to the legal domain through three main strategies, namely fine-tuning, further pre-training, or from-scratch pre-training of a BERT model on a legal corpus (e.g., [20, 115]). The language modeling component of LawNet-Viz is made flexible to the specific configuration of a legal BERT-based model, allowing for different extensions backed by TensorFlow or PyTorch libraries.

Node and edge features LawNet-Viz is designed to handle different features associated with articles and article references.

Each node (i.e., article) pair ⟨source, target⟩ is assigned a real-valued weight, which corresponds to the cosine similarity between the vectors of the two articles, according to a selected representation model (cf. Language modeling paragraph).

Attributes at node level include both topological/network-layout information and properties of the corresponding article, such as article ID, article title, article text, book ID, cluster ID, layout coordinates, color, size, number of references (overall, incoming, outgoing), and various centrality scores. The latter include betweenness, closeness, PageRank, Eigenvector centrality, and two variants of TextRank, a well-known weighted extension of the PageRank, where edges are weighted according to some similarity measure; the two TextRank variants implemented in LawNet-Viz utilize the edge features we discussed above. Note that besides the microscopic-level node attributes, the 'book ID' and 'cluster ID' attributes allow for exploring the network at two different mesoscopic levels, i.e., the one corresponding to the logical structure of the law code, and the other one corresponding to the outcome of some graph clustering method (e.g., community memberships). Also, the size and color attributes of any node are defined according to the article's centrality and membership, respectively; by default, the size is set proportionally to the node degree and the color code matches the cluster ID if available, otherwise the book ID.

[2] https://huggingface.co/docs/transformers/index

Data format All data in LawNet-Viz are structured into a multi-line JSON format enclosing nodes, edges, and their corresponding attributes as discussed above.

We emphasize that our JSON data structure can easily be produced through most of the existing frameworks for network analysis, thus ensuring broad compatibility and interoperability with LawNet-Viz. As a case in point, our JSON data complies with the one that can be exported from Gephi [8] (through its *JSON Exporter* or *SigmaExporter* plugins), which is nowadays widely recognized as the de-facto standard tool for general-purpose network exploration and analysis. Consequently, such a Gephi-compatibility allows users to abstract from programming needs and straightforwardly build a JSON object — along with a set of topological statistics — to be used within our platform; this also includes the capability to transfer the network layout and color coding directly to LawNet-Viz. Nonetheless, we point out that the JSON data structure can also be extended by users to include additional node and edge features.

Web application Our choice in the implementation of the LawNet-Viz web platform was to leverage exclusively on offline or server-side data processing to minimize potential client-side loads, and make the platform suitable for any desktop or mobile device.

As for the front-end part, we utilize the widely known, free and open-source frameworks *Bootstrap* (v. 5.1.3) [3] and *DataTables* (v. 1.11.4), [4] where the former effectively supports the development of responsive, visually pleasant web user interfaces, and the latter enables the interactive visualization of data tables.

The back-end part is created upon the open-source web-based network visualization *vis.js* library (v. 9.1.0). [5] Note however that, given the visualization and interaction requirements of LawNet-Viz, we extended the core APIs of the "vanilla" version of *vis.js*, and developed specific Python3 scripts, in order to come up with a new set of operations to cover aspects that ensure seamless interplay between topological and textual features, such as dynamic interaction with the articles' node and their texts, responsive filtering of the network based on thresholding of the strength of the article links (i.e., similarity between an article and a referred one), access to different text representation models.

[3] https://getbootstrap.com/
[4] https://datatables.net/
[5] https://visjs.org/

Chapter 6

TrustSearch: NLP to promote critical thinking in search engine.

This chapter is based on the research I carried out during the PhD visiting period abroad at (UPV) University Polytechnic of Valencia, Spain. During this period I had the honor to collaborate with the multicultural and multidisciplinary academic community at PRHLT(Pattern Recognition and Human Language Technology Research Center) of the UPV. I actively participated in the TrustSearch project, a research project funded by the European Union, developing an AI tool that offers a new search experience to promote critical thinking, as described in Section6.1. With a focus on the stance detection task, I collaborated in the design, development, and project management phases and made a significant contribution to the creation and processing of the dataset 6.2, as well as the development of an innovative approach of Natural Language Processing (NLP) techniques for the stance detection task 6.3. Specifically, I utilized Transformer encoder-decoder based model initially developed for entailment tasks and adapted them for performing stance detection. Additionally, i employed Language Models (LLMs) with prompt engineering techniques. Finally, I developed a working demo 6.4 of the project.

6.1 Proposed Approach

Brief description. In the last decade we appreciated an increasing mistrust in all information sources. Following the Edelman Trust Barometer 2021[1], we see that trust in information is at record lows. Also, confidence in search engines (-9%) and traditional media (-12%) has decreased in the last two years. According to this, the Next generation of the Internet should

[1] https://www.edelman.com/trust/archive

comprise a change in the way we experience search and discover data and resources on the internet and web. A new generation of search engines that promotes critical thinking must be developed. Our project aims to design an AI tool that, on the basis of a user query on a controversial issue, shows a plural media landscape where the user has the possibility of; (1) analyzing the stance of different media; (2) distinguishing the sources in terms of audience (number of visits) and (3) leading political ideology.

Scope: The mistrust problem. The mistrust in information sources could be linked to the proliferation of fake news and other disinformation practices, but we think that a key factor is also that the audience perceives ideological bias in the media. Again, following Edelman Trust Barometer 2021, we see that 59% of people agree with the idea that "most news organizations are more concerned with supporting an ideology or political position than with informing the public" and 61% think that the media are not doing well at being objective and non-partisan. In some European countries such as Spain (73%) or Italy (75%) the percentage of people that support this belief is even greater. Such penetration of distrust in audiences means that the problem we face goes beyond content verification and relies on polarization and ideological bias affecting information sources. Three complementary phenomena could play a role in this process of increasing mistrust: (1) polarization-based narratives have become pervasive; (2) search engines have equalized information sources by presenting the most serious and neutral journalism and the most propagandistic news sites on the same level; and (3) algorithms facilitate a poor information diet exacerbating the "echo chamber effect"[2]. This project focuses on factor 2, trying to offer a new search experience that generates greater confidence in the search itself. There is a big room for improvement because following Edelman Trust Barometer 2021 only 29% of people have good information hygiene practices, that is: (1) regular engagement with news; (2) engagement with differing points of view, avoiding echo chambers; (3) information verification activities; (4) prevention of spreading disinformation. According to this, we need to go beyond the identification of disinformation and develop an ambitious strategy that promotes critical thinking and combines the principles of Media Literacy with new tools of Artificial Intelligence. As the European Council recommends, we need to work on solutions specifically aimed at minimizing the impact of filter bubbles, diversifying exposure to different views.

Ambition: To promote critical thinking and go beyond static, one-dimensional search results. To encourage critical thinking, we need to develop AI tools considering the perception of the media as ideologically biased. Several websites give information about the ideological bias of the source, although this activity is not linked to a particular subject from a user query. Our project aims to create an AI tool showing the different approaches of

[2]Pariser, E.(2011) The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.

media to a controversial issue. Given a user query about a controversial topic, the tool will: (1) find news on the same topic in news sources that are in different ideological positions; (2) present to the users an interactive map that situates the news attending their stance on the topic (strongly in favor, in favor, neutral/none, against, strongly against); and (3) give information about the source: leading political ideology and his audience (number of visits). Our project aims to create an AI tool showing an interactive map the different approach of the media to a particular event from one user query. The AI tool will promote critical thinking, encouraging users to be exposed to news of different political orientation and lines of thought. The research will examine human computer interaction to evaluate the effect of being exposed to a plurality of news on reducing polarization.

Objectives. General Objective: To integrate the results of research in stance detection, media bias and news classification in a new search tool that aims to promote critical thinking and hygiene information practices in web searches, through an innovative way of structuring, analyzing and visually representing online news around a given topic. Specific Objectives:

- Develop the TrustSearch tool to analyze a controversial topic of a given user query and find a plurality of news about it, giving the user information about the stance of each media (strongly in favor, in favor, neutral/none, against, strongly against);
- Develop a visual representation of these results in an interactive map that allows users to explore the different media approaches to the issue in terms of: (1) stance, (2) leading political ideology of the source, (3) audience of the source (number of visits).

Innovation: Beyond the state of the art To cope with the phenomenon defined as information pollution at a global scale[3], AI researchers have investigated different ways to improve the exposure to divergent sources of information by: (1) stance detection, (2) detecting media bias, (3) employing algorithms for content selection, and (4) changing the depiction of information to encourage users to be exposed to a plurality of sources and opinions. In this chapter we focus on the stance detection.

Stance detection. Stance detection is an emerging opinion mining paradigm for various social and political applications. Recently there has been a growing research interest in detecting automatically the stance from opinionated texts, especially if about controversial topics (e.g., vaccines, climate change, etc.). The recent survey of ALDayel and Magdy (2021) [2] presents an exhaustive review of stance detection techniques, the different types of targets, the features set used, and the best performing machine learning / deep learning approaches.

[3]Wardle, C. and Derakhshan, H. (2017). INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making Information Disorder. Council of Europe Report DGI (2017)09

6.2 Data

In order to validate the proposed stance detection approaches, we select three topics: climate change, immigration and COVID vaccination from 30 news outlets (15 in Spanish and 15 in English). We use two different approaches in order to create a pilot dataset: a RSS topic-based approach and a Query-based approach. Figure 6.1 shows the dataset processing pipeline. It should be noted, that the same processing pipeline was followed for both dataset generation strategies, with the only difference being that the scoring algorithm stage was skipped in the case of the query-based dataset since the articles were already relevant to the topics and no supplementary filtering was needed. All the articles were collected and processed by a proprietary web scraper. The purpose of the web scraper is to extract any valuable information and metadata from an article and index it in an internal search engine/Apache Solr[4]. The next stage in the pipeline was the execution of the scoring algorithm which is a basic filtering mechanism for keeping the relevant articles 1from the dataset. We applied the powerful lemmatization and stemming of the Apache Solr so that we could perform a keyword-based scoring procedure which utilizes the stem of the word and the context in which the word is being used. The scoring algorithm and its visual representation is shown in Figure 6.2. We loop over all articles for each available keyword of the three topics. For each topic, an article receives one point if a keyword was found at least once in its content. The more points an article receives, the more relevant to this category it is considered. The final stage of the pipeline was the cleaning process. This stage consists of procedures which are executed step-by-step in a sequential order. These are the steps:

- duplicated links resolution;
- remove instances with missing or too short “title” (it indicates potential scraping problems);
- remove instances with missing or too short “headline” (it indicates potential scraping problems);
- remove instances with missing or too short “text” (it indicates potential scraping problems);
- remove instances with duplicated text, in particular where there is some overlap between chunks of text (it indicates potential scraping problems, such as the text being just part of static page);

[4]https://solr.apache.org/

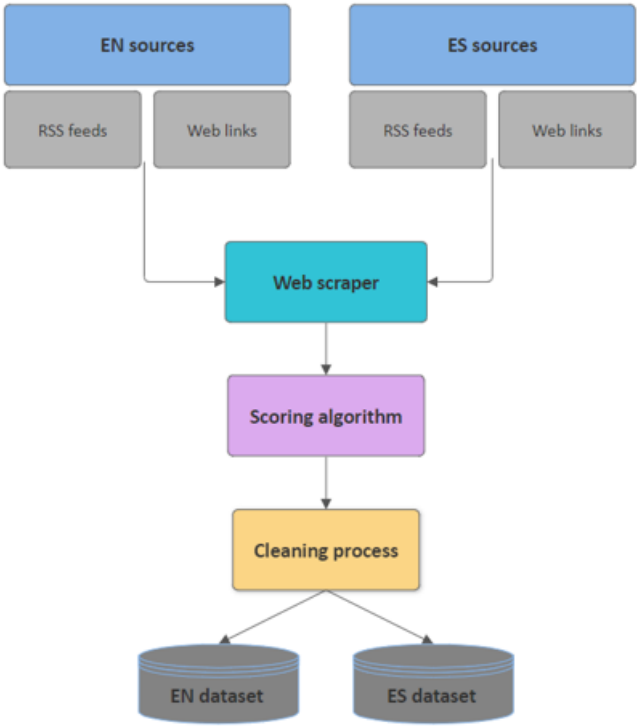


Fig. 6.1 The dataset processing pipeline.

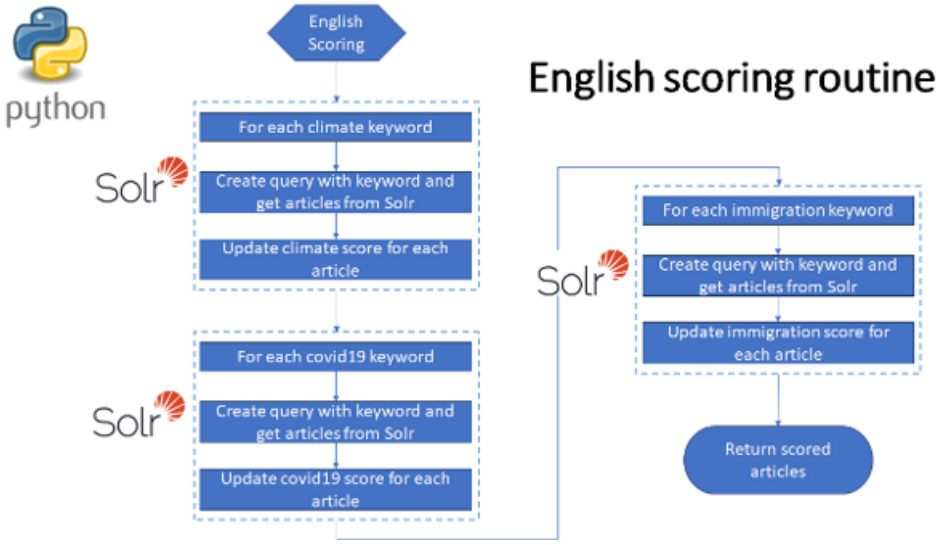


Fig. 6.2 Scoring algorithm routine.

- summarization of article text, even in order to remove noise from the text.

6.2.1 RSS topic-based

The RSS-topic based approach in creating the dataset was to aggregate the news articles from the curated list of news outlets in Spain and the UK through their RSS feeds. From this initial dataset, we wanted to extract only the articles that fall into one of the three major topics, and score them according to their relevance to the topic. A list of keywords per topic was created for the above scoring procedure. The list of keywords consists of words that are strongly related to each topic.

6.2.2 Query-based

We found that, using the RSS-topic based approach, many instances were not properly related to the topic. Therefore, we decided to address the creation of the pilot dataset with a second strategy based on user queries. For this second approach, we developed a pilot study with six users. Users were asked two questions about each selected topic:

- Please think of a question about (climate change/immigration/COVID-19 vaccination) that you would ask a search engine like Google, or similar, and write it below.
- What is your personal opinion on this question? Do you have any previous ideas that lead you to ask it?

From this pilot study, we develop a dataset based on the following questions in English and in Spanish as well. The English version:

- climate change
  - Is the reversal of the effects of climate change possible?
  - What causes climate change?
  - How can you contribute to the user level to combat climate change?
- covid vaccination
  - What side effects really have vaccines that we are not told?
  - Was the vaccine useful? Does the vaccine really save lives?
  - Should vaccines against diseases such as Covid 19 be free?Is the reversal of the effects of climate change possible?
- immigration

  - Is immigration linked to crime?
  - Are immigrants receiving more help from the state than national people?
  - How to solve the problem of immigration in Europe?

The Spanish version:

- climate change
  - ¿A qué se debe el cambio climático?
  - ¿Cómo se puede contribuir a nivel de usuario para combatir el cambio climático?
  - ¿Es posible la reversión de los efectos del cambio climático?
- covid vaccination
  - ¿Qué efectos secundarios tiene realmente las vacunas que no nos cuentan?
  - ¿Fue la vacuna útil? ¿Realmente salvo vidas la vacuna?
  - ¿Deberían ser gratuitas las vacunas de la enfermedad llamada COVID 19?
- immigration
  - ¿La inmigración está relacionada con la delincuencia?
  - ¿Como solucionar el problema de la inmmigración en Europa?
  - ¿Reciben los inmigrantes más ayudas que los nacionales en los países europeos?

Based on the above queries and on a fixed list of media sources, we collect the news from Google News. To deal with this, we use python and some libraries (e.g., GoogleNews[5]) crawling through the search results and extracting valuable links while ensuring they adhere to specific criteria, such as being news articles and originating from a predefined list of media sources. In particular, we use:

- &tbm=nws parameter to filter the search results to include only news articles;
- &as_sitesearch= parameter to limit the results to a specific list of media sources;
- pagination, to collect a comprehensive set of news articles;
- information about publishing date, for example, to filter out articles published before the coronavirus pandemic.

[5]https://pypi.org/project/GoogleNews/

Therefore, we use libraries like BeautifulSoup[6] for HTML parsing. At the end, for this dataset we have the following information: [id, link, query, title, headline, text, summary, media, date, quality, quantity, diversity, rarity, score].

6.3 Stance detection

In order to face the stance detection problem of news media articles regarding the topics of interest, we experiment with two different approaches. The first one named as Definition-based approach is based on a fixed definition of the topic. The latter named as User-driven approach is based on the user's previous ideas about a topic.

6.3.1 Definition-based approach

This approach relies on a predefined definitions associated with a specific topic checking whether these definitions are present in the article. Definition-based stance detection approach is designed to provide the estimation of the general idea of an article about a chosen topic. There are three possible answers about the stance of an article:

1. article is *in favor* w.r.t. the general idea about the topic;
2. article is *neutral* w.r.t. the general idea about the topic;
3. article is *against* w.r.t. the general idea about the topic.

This approach is based on prompt engineering and utilizes large language models (LLMs) as zero-shot classification estimators. According Zhang et al. [112] and our experiments, LLMs are quite capable in solving this task, and their performance is currently the state of the art. However, there are advantages and disadvantages to using LLMs. The upside is that LLMs can produce good performance with explanations without needing a specific dataset and training for each topic. The downside is that using LLMs is more expensive than using smaller models. There are also issues specific to commercial LLMs. These models are opaque, i.e., the model parameters are unknown. There is also a risk of continuity of service, the company can stop providing access to the model in future. For experiments within the project we use two LLMs, GPT-3.5-Turbo and GPT-4, and three different topics: 1) climate change, 2) vaccination strategy against covid-19 and 3) immigration. First we checked if

[6]https://pypi.org/project/beautifulsoup4/

LLMs (GPT-3.5[7] and GPT-4[8]) have a good notion about these topics. Thus, we checked the quality of definition generated by LLMs. These are the definitions generated by GPT-4.

> Climate change: *"Climate change refers to long-term shifts in temperature patterns and weather conditions on Earth. It is primarily caused by human activities such as the burning of fossil fuels, deforestation, and industrial processes, which release greenhouse gases into the atmosphere. These gases, like carbon dioxide, trap heat from the sun and result in a warming effect on the planet. Climate change leads to various effects, including rising sea levels, extreme weather events, changes in ecosystems, and threats to human health and livelihoods. Taking action to reduce greenhouse gas emissions and adapt to these changes is crucial in addressing climate change."*
>
> Covid vaccination: *"The vaccination strategy against COVID-19 involves globally distributing and administering approved vaccines to prioritize high-risk groups such as healthcare workers, the elderly, and people with underlying health conditions. This strategy aims to reduce transmission, severe illnesses, and deaths caused by the virus. Vaccination campaigns usually follow a phased approach, starting with priority groups and gradually expanding to the general population. Additionally, strategies include ensuring vaccine supply, establishing vaccination centers, implementing public education campaigns, and monitoring vaccine safety and efficacy."*
>
> Immigration: *"Immigration refers to the process by which individuals enter and settle in a foreign country with the intention of residing there permanently or for an extended period. It involves crossing international borders and often requires complying with legal requirements, such as obtaining visas or work permits. Immigration can be motivated by various factors, including economic opportunities, seeking asylum or refuge, family reunification, or pursuing education and better living conditions."*

The definitions given by LLMs were consistent with the human generated ones. After that, we generated by LLMs the arguments in favor and against these topics. Both LLMs generated arguments similar to those generated by humans. The final step was to verify that the models were providing well-formed responses so that the model prediction could be extracted. The both LLMs provide well-formatted results, making them suitable for solving this task in zero-shot manner. In performance experiments, we query these LLMs with the title, subtitle, and abstract of the article body. After that, the model is asked to evaluate the stance of the article according to the topic. To explore the reasoning behind model predictions, we expand the query and make these models provide explanations to elaborate their predictions. The prediction of models were verified by humans and in most cases the estimation of LLMs were consistent with humans. In addition, we further extended classification approach to estimate intensity of the articles' stance, i.e., how much the article is against or in favor of the topic. For this reason, we did the stance detection on the sentence level, for each sentence in text separately. We calculate the final score as a weighted sum of the sentence stance scores, where sentence length is used to calculate weight/importance of the sentence. The final prediction result is a continuous value between -1 and 1, with a value of -1 representing an attitude completely against and a value 1 representing an attitude completely in favor of the topic. In our experiments, both LLMs achieve good results for sentence-level stance detection, however the GPT-4 model allows for more versatile approaches.

[7] https://platform.openai.com/docs/models/gpt-3-5
[8] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

6.3.2 User-driven approach

The user-driven stance detection approach, rooted in entailment task, diverges from the definition-based stance detection methodology. Unlike the latter, which aims to gauge the general population's stance on a chosen topic, the user-driven approach centers around a real user query and its previous personal idea about (i.e., the stance). In this context, the goal is to determine the entailment between two texts. In particular, given two sentences, named *premise* and *hypothesis* respectively, the goal of the entailment task is to identify whether the hypothesis is true based on the information given in the premise. If we call the newspaper article text as premise and the user's pair <query,stance> as hypothesis we will be able to obtain in output an answer to the following question: "Is the user's personal stance true (or entailed, or we can infer this) based on the information contained in the article?". In this regard, we have used an encoder-decoder Transformer-based model named BART [53] since the promising performance on the entailment tasks. In particular, we have used a checkpoint for bart-large-mnli[9] after being trained on the MultiNLI (MNLI)[10] dataset, then we have applied a NLI-based Zero Shot Text Classification between two classes: the user's stance and opposite of the user's stance. The input consists basically in the user's question and the user's stance. Indeed, the topic is not mandatory for this approach and the opposite of the user's stance can be automatically generated. The output consists of a binary classification between the aforementioned two classes. In this way we are able to obtain a clear view of the relation between the user's stance and the news article's stance promoting the critical thinking of the user. In contrast to traditional methods, this approach empowers users to actively guide the model's understanding, making it particularly adaptable to diverse contexts and individual perspectives. The emphasis on user-driven input allows for a more nuanced and tailored stance detection process. It should be noted that in this way we obtain a topic-independent approach which is totally generalized and useful for a real case scenario.

6.4 Demo

In this section we show some screenshots from the demo developed to show the capabilities of the proposed tool.

Figure 6.3 shows the user interface for the Definition-based approach. It should be noted that the user can choose between the three topics selected for the pilot study (i.e., climate change, covid vaccination and immigration) through radio buttons. When the user has selected a topic, then the topic definition, generated by GPT-4 will appear. A this point,

[9]https://huggingface.co/facebook/bart-large-mnli
[10]https://cims.nyu.edu/ sbowman/multinli/

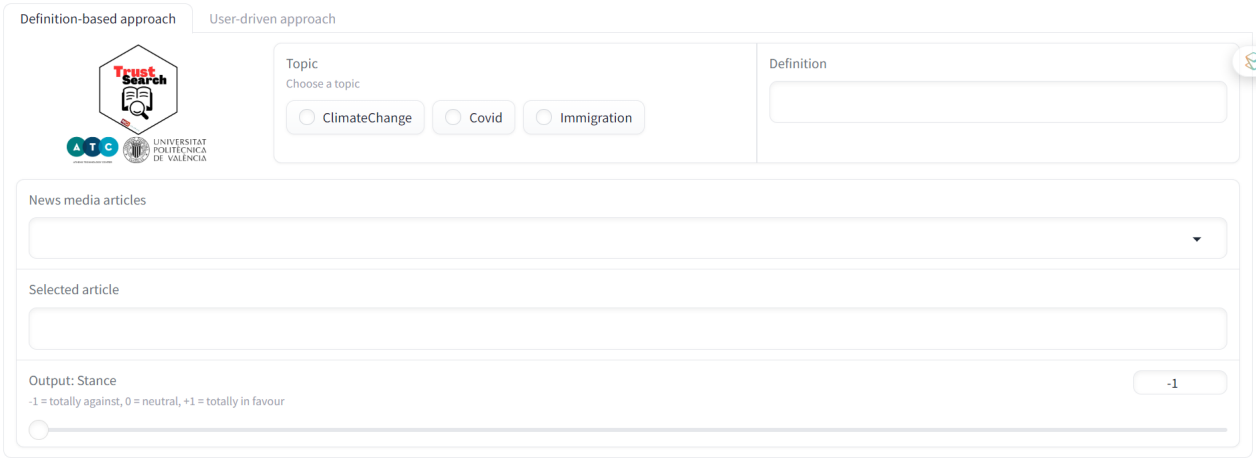


Fig. 6.3 Input fields for the Definition-based approach.

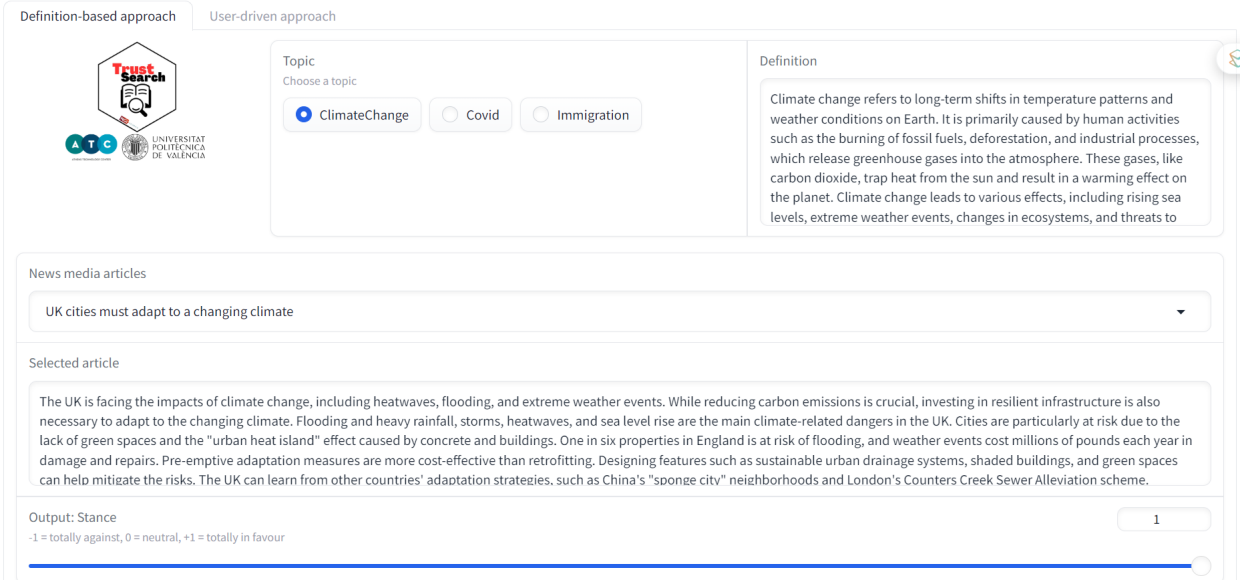


Fig. 6.4 Example for the Definition-based approach.

the user can pick one news media article related to the selected topic. The output stance will be visible at the bottom in Figure 6.4.

Figure 6.5 shows the user interface for the User-driven approach. The user interface is a little bit different in that case because the input data are different. In fact, here the user can express a question and the personal previous idea about the question itself(i.e., user's stance). For experimental reasons we decided to generate the opposite of the user's stance as well to facilitate understanding of the model's output. In Figure 6.6, it should be noted which the system correctly declare "Excess emissions cause climate change" as the most probable class with a confidence of 98%.

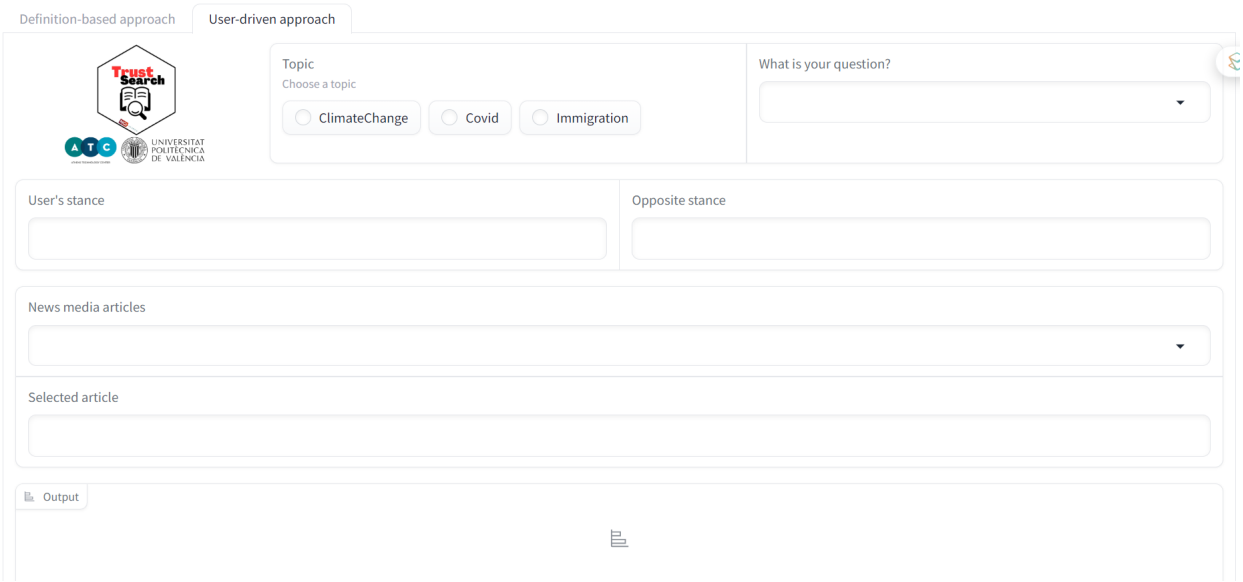


Fig. 6.5 Input fields for the User-driven approach.

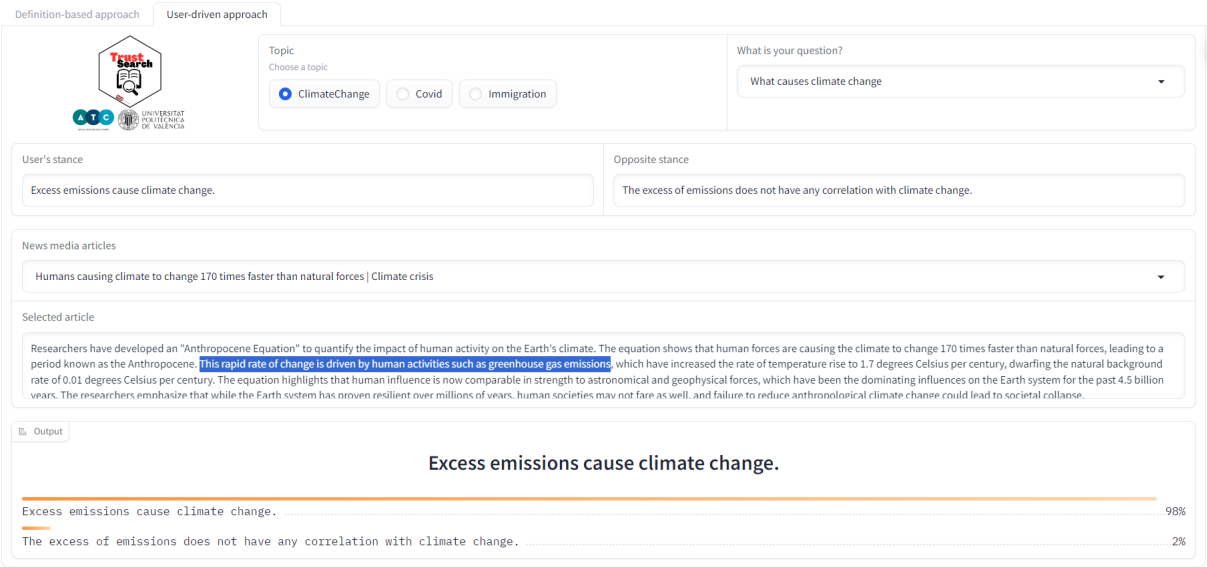


Fig. 6.6 Example of output for the User-driven approach. It should be noted which the system correctly declare “Excess emissions cause climate change” as the most probable class with a confidence of 98%.

Chapter 7

Conclusions

Summary. In this PhD thesis, we presented LamBERTa, a novel BERT-based language understanding framework for law article retrieval as a prediction task. One key aspect of LamBERTa is that it is designed to deal with a challenging learning scenario, where the multi-class classification setting is characterized by hundreds of classes and very few, per-class training instances that are generated in an unsupervised fashion. We looked into the effects of domain-adaptation in combination with the task-adaptation to learn Italian BERT models (LamBERTa) for the task of ICC article prediction. To this purpose, we investigated the role of a recently defined Italian legal BERT into our framework, as well as the effects of enhancing the tokenizer with new terms selected from the target legal corpus, by varying the setting for the initial token embeddings of such terms. We expect that our work can pave the way for further exploration of domain-/task-adaptation for (Italian) legal BERT models.

We explored the explainability of our LamBERTa models through two distinct approaches. Firstly, using model-level explainability which delves into comprehending how these models establish intricate relationships among textual tokens. Secondly, through instance-level explainability employs LIME to elucidate the behavior of the underlying classifier in the vicinity of a query instance.

Note also that, while focusing on the Italian Civil Code in first version, the LamBERTa architecture can easily be generalized to learn from other law code corpus as the GDPR. Therefore, we addressed the problem of GDPR article retrieval through the use of PLMs, and in particular LamBERTa architectures which include both domain-general and domain-specific pre-trained BERT models, further powered by self-supervised task-adaptive pre-training stages, with or without data enrichment based on recitals. To the best of our knowledge, this is the first study that explores a number of aspects concerning both domain-adaptive and task-adaptive legal pre-trained language models for the task of GDPR article retrieval.

Furthermore, in this work, by taking a network analysis and mining perspective, we presented the first study of citation networks that can be inferred from the ICC articles. Our analysis of the structural features of such networks has shed light on valuable hidden patterns, such as the linkage between community memberships of articles and their topical structure, paving the way for new interpretations and study of the ICC. It is also worth noticing that our proposed methodology can easily be generalized to other civil law code systems presenting a similar organization as the ICC, i.e., developed upon a logical structure of the corpus into books, and their internal subdivisions. It is our opinion that exploring the network underlying the law reference relations as well as semantic affinity among articles in a law code is highly important to support a variety of tasks for legal information processing and understanding, including statute law retrieval, entailment and question answering. In this respect, we developed LawNet-Viz, a web-based tool for the modeling, analysis and visualization of law reference networks extracted from a statute law corpus. The purpose of our research is to show that a deep-learning-based civil-law article retrieval method can be helpful not only to legal experts to reduce their workload, but also to citizens who can benefit from such a system to save their search and consultation needs.

Finally, we presented the ambitious goal of the TrustSearch project to develop an AI tool that offers a new search experience to promote critical thinking. In order to face the arising challenges, we developed two stance detection approaches (i)Definition-based approach (ii)User-driven approach, which leverage on powerful LLMs (e.g., ChatGPT) and on an encoder-decoder Transformer-based model (e.g., BART) previously trained on the entailment task. In this regard, we shown the effectiveness of the AI to propose a change in the way we experience search and discover data and resources.

We believe these steps will contribute to the development of powerful tools that can deliver unprecedented solutions to different problems.

Ongoing work. We are currently working on an evaluation of our proposed models on the CMS.Law GDPR Enforcement Tracker database. Preliminary experimental results have shown the effectiveness of our models, both in absolute terms and in relation to the article classification approach proposed in [3] relying on labeled LDA and an LSTM model (cf. Related Work).

We also plan to define effective methods to integrate task-oriented knowledge on a pre-trained (domain-adaptive) model, as well as a from-scratch pre-trained Italian legal BERT model.

As previously discussed, our models for GDPR article retrieval can serve as a basis for a variety of similarity based tasks, including question-answering. Indeed, we are working on further developments of our models to deal with GDPR compliance and violation checking

tasks. In particular, we are developing a hybrid framework based on a combination of BERT-like models and ChatGPT-like models in order to take advantage of similarity search and classification capabilities of the former and conversational functionality of the latter.

References

[1] Agnoloni, T. and Pagallo, U. (2015). The Case Law of the Italian Constitutional Court, Its Power Laws, and the Web of Scholarly Opinions. In Proc. of the 15th International Conference on Artificial Intelligence and Law, pages 151–155. Association for Computing Machinery.

[2] ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. Information Processing amp; Management, 58(4):102597.

[3] Aleroud, A., Masalha, F., and Saifan, A. A. (2021). Identifying GDPR privacy violations using an augmented LSTM: toward an ai-based violation alert systems. In Proc. of the IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pages 1617–1624. IEEE.

[4] Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., and Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Computer Science, 2:e93.

[5] Amaral, O., Azeem, M. I., Abualhaija, S., and Briand, L. C. (2022). NLP-based Automated Compliance Checking of Data Processing Agreements against GDPR. CoRR, abs/2209.09722.

[6] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. (2023). Palm 2 technical report.

[7] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proc. ICLR.

[8] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. Proceedings of the International AAAI Conference on Web and Social Media, 3(1):361–362.

[9] Bengio, S., Dembczynski, K., Joachims, T., Kloft, M., and Varma, M. (2019). Extreme classification. Technical report, Report from Dagstuhl Seminar 18291.

[10] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10:P10008.

[11] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics Report, 424:175–308.

[12] Boella, G., Caro, L. D., and Humphreys, L. (2011). Using classification to support legal knowledge engineers in the Eunomos legal document management system. In Proc. Workshop on Juris-informatics (JURISIN).

[13] Boniol, P., Panagopoulos, G., Xypolopoulos, C., Hamdani, R. E., Amariles, D. R., and Vazirgiannis, M. (2020). Performance in the Courtroom: Automated Processing and Visualization of Appeal Court Decisions in France. In Proc. of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, volume 2645 of CEUR Workshop Proceedings, pages 11–17. CEUR-WS.org.

[14] Branting, K., Weiss, B., Brown, B., Pfeifer, C., Chakraborty, A., Ferro, L., Pfaff, M., and Yeh, A. S. (2019). Semi-supervised methods for explainable legal prediction. In Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), pages 22–31.

[15] Branting, L. K., Yeh, A. S., Weiss, B., Merkhofer, E. M., and Brown, B. (2017). Inducing predictive models for decision support in administrative adjudication. In AI Approaches to the Complexity of Legal Systems - AICOL Workshops 2015-2017, volume 10791, pages 465–477.

[16] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[17] Chalkidis, I., Androutsopoulos, I., and Aletras, N. (2019a). Neural legal judgment prediction in english. In Proc. ACL, pages 4317–4323. Association for Computational Linguistics.

[18] Chalkidis, I., Androutsopoulos, I., and Michos, A. (2018). Obligation and Prohibition Extraction Using Hierarchical RNNs. In Proc. Annual Meeting of the Association for Computational Linguistics (ACL), pages 254–259.

[19] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019b). Extreme Multi-Label Legal Text Classification: A case study in EU Legislation. In Proc. Natural Legal Language Processing (NLLP) Workshop of NAACL-HLT, pages 78—-87.

[20] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: the muppets straight out of law school. CoRR, abs/2010.02559.

[21] Chalkidis, I. and Kampas, D. (2019). Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif. Intell. Law, 27(2):171–198.

[22] Conrad, J. G. and Branting, L. K. (2018). Introduction to the special issue on legal text analytics. Artif. Intell. Law, 26(2):99–102.

[23] Crotti Junior, A., Orlandi, F., Graux, D., Hossari, M., O'Sullivan, D., Hartz, C., and Dirschl, C. (2020). Knowledge Graph-Based Legal Search over German Court Cases. In Proc. of the Semantic Web Conference: ESWC 2020 Satellite Events, pages 293–297.

[24] Dadgostari, F., Guim, M., Beling, P., Livermore, M. A., and Rockmore, D. (2021). Modeling law search as prediction. Artif. Intell. Law, 29(1):3–34.

[25] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT, pages 4171–4186.

[26] Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Mach. Learn., 42(1/2):143–175.

[27] Do, P., Nguyen, H., Tran, C., Nguyen, M., and Nguyen, M. (2017). Legal question answering using ranking SVM and deep convolutional neural network. CoRR, abs/1703.05320.

[28] Du, C. and Huang, L. (2018). Text classification research with attention-based recurrent neural networks. Int. J. Comput. Commun. Control, 13(1):50–61.

[29] Elluri, L., Chukkapalli, S. S. L., Joshi, K. P., Finin, T., and Joshi, A. (2021). A BERT based approach to measure web services policies compliance with GDPR. IEEE Access, 9:148004–148016.

[30] Filtz, E. (2017). Building and processing a knowledge-graph for legal data. In Proc. of the 16th International Semantic Web Conference, pages 184–194. Springer International Publishing.

[31] Filtz, E., Kirrane, S., and Polleres, A. (2021). The linked legal data landscape: linking legal data across different countries. Artificial Intelligence and Law, 29(4):485–539.

[32] Fowler, J. H., Johnson, T. R., Spriggs, J. F., Jeon, S., and Wahlbeck, P. J. (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis, 15(3):324–346.

[33] Gan, L., Kuang, K., Yang, Y., and Wu, F. (2021). Judgment prediction via injecting legal knowledge into neural networks. In Proc. AAAI, pages 12866–12874. AAAI Press.

[34] Goldberg, Y. (2017). Neural network methods in natural language processing. Morgan and Claypool Publishers.

[35] Goodfellow, I., Y, Y. B., and Courville, A. (2016). Deep learning. MIT Press, Cambridge.

[36] Greco, C. M., Simeri, A., Tagarelli, A., and Zumpano, E. (2023). Transformer-based language models for mental health issues: A survey. Pattern Recognition Letters, 167:204–211.

[37] Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 8342–8360. ACL.

[38] Hacker, P., Krestel, R., Grundmann, S., and Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. Artif. Intell. Law, 28(4):415–439.

[39] Hamdani, R. E., Mustapha, M., Amariles, D. R., Troussel, A. C., Meeùs, S., and Krasnashchok, K. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. In Proc. of the Eighteenth International Conference for Artificial Intelligence and Law (ICAIL 2021), pages 40–49. ACM.

[40] Hu, Z., Li, X., Tu, C., Liu, Z., and Sun, M. (2018). Few-shot charge prediction with discriminative legal attributes. In Proc. COLING, pages 487–498. Association for Computational Linguistics.

[41] Jain, A. K. and Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall.

[42] Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. J. Documentation, 60(5):493–502.

[43] Junior, A. C., Orlandi, F., O'Sullivan, D., Dirschl, C., and Reul, Q. (2019). Using Mapping Languages for Building Legal Knowledge Graphs from XML files. In Proc. of the Blockchain enabled Semantic Web Workshop (BlockSW) and Contextualized Knowledge Graphs (CKG) Workshop co-located with the 18th International Semantic Web Conference, volume 2599 of CEUR Workshop Proceedings. CEUR-WS.org.

[44] Kim, M., Xu, Y., and Goebel, R. (2015). A convolutional neural network in legal question answering. In Proc. Int. Workshop on Juris-informatics JURISIN.

[45] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proc. EMNLP, pages 1746—-1751.

[46] Koniaris, M., Anagnostopoulos, I., and Vassiliou, Y. (2017). Network analysis in the legal domain: a complex model for European Union legal sources. Journal of Complex Networks, 6(2):243–268.

[47] La Cava, L., Simeri, A., and Tagarelli, A. (2021). The Italian civil code network analysis. In Proc. of the Relations in the Legal Domain Workshop co-located with the 15th International Conference on Artificial Intelligence and Law, volume 2896, pages 3–16.

[48] La Cava, L., Simeri, A., and Tagarelli, A. (2022). Lawnet-viz: A web-based system to visually explore networks of law article references. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, page 3300–3305, New York, NY, USA. Association for Computing Machinery.

[49] Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In Proc. AAAI, pages 2267—2273.

[50] Lee, B., Lee, K.-M., and Yang, J.-S. (2019). Network structure reveals patterns of legal complexity in human society: The case of the constitutional legal network. PLOS ONE, 14(1):1–15.

[51] Lettieri, N., Altamura, A., Faggiano, A., and Malandrino, D. (2016). A computational approach for the experimental study of eu case law: analysis and implementation. Social Network Analysis and Mining, 6(1):56.

[52] Lettieri, N., Altamura, A., and Malandrino, D. (2017). The legal macroscope: Experimenting with visual legal analytics. Information Visualization, 16(4):332–345.

[53] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

[54] Li, L., Bi, Z., Ye, H., Deng, S., Chen, H., and Tou, H. (2021). Text-guided legal knowledge graph reasoning. In Proc. of the China Conference on Knowledge Graph and Semantic Computing, pages 27–39.

[55] Li, S., Zhang, H., Ye, L., Guo, X., and Fang, B. (2019). MANN: A multichannel attentive neural network for legal judgment prediction. IEEE Access, 7:151144–151155.

[56] Licari, D. and Comandè, G. (2022). ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In Proc. KM4LAW Workshop with the 23rd Int. Conf on Knowledge Engineering and Knowledge Management, volume 3256 of CEUR Workshop Proceedings. CEUR.

[57] Lin, W., Kuo, T., Chang, T., Yen, C., Chen, C., and Lin, S. (2012). Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. IJCLCLP, 17(4).

[58] Liu, C. and Hsieh, C. (2006). Exploring phrase-based classification of judgment documents for criminal charges in chinese. In Proc. ISMIS, pages 681—-690.

[59] Liu, P., Qiu, X., and Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In Proc. IJCAI, pages 2873—-2879.

[60] Liu, Y.-H., Chen, Y.-L., and Ho, W.-L. (2015). Predicting associated statutes for legal problems. Information Processing & Management, 51(1):194–211.

[61] Long, S., Tu, C., Liu, Z., and Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In Proc. Chinese Computational Linguistics, volume 11856 of Lecture Notes in Computer Science, pages 558–572.

[62] Lorè, F., Basile, P., Appice, A., de Gemmis, M., Malerba, D., and Semeraro, G. (2023). An AI framework to support decisions on GDPR compliance. Journal of Intelligent Information Systems, pages 1–28.

[63] Luo, B., Feng, Y., Xu, J., Zhang, X., and Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In Proc. EMNLP, pages 2727—-2736.

[64] Mazzega, P., Bourcier, D., and Boulet, R. (2009). The Network of French Legal Codes. In Proc. of the 12th International Conference on Artificial Intelligence and Law, ICAIL '09, page 236–237.

[65] Medvedeva, M., Vols, M., and Wieling, M. (2020). Using machine learning to predict decisions of the european court of human rights. Artif. Intell. Law, 28(2):237–266.

[66] Moodley, K., Hernandez-Serrano, P. V., Zaveri, A. J., Schaper, M. G., Dumontier, M., and Van Dijck, G. (2020). The case for a linked data research engine for legal scholars. European Journal of Risk Regulation, 11(1):70–93.

[67] Morimoto, A., Kubo, D., Sato, M., Shindo, H., and Matsumoto, Y. (2017). Legal question answering system using neural attention. In Proc. 4th ICAIL Competition on Legal Information Extraction and Entailment (COLIEE), volume 47 of EPiC Series in Computing, pages 79–89.

[68] Moser., M. and Strembeck., M. (2019). An Analysis of Three Legal Citation Networks Derived from Austrian Supreme Court Decisions. In Proc. of the 4th International Conference on Complexity, Future Information Systems and Risk, pages 85–92.

[69] Nallapati, R. and Manning, C. D. (2008). Legal Docket Classification: Where Machine Learning Stumbles. In Proc. EMNLP, pages 438–446.

[70] Nanda, R., Adebayo, K. J., Caro, L. D., Boella, G., and Robaldo, L. (2017). Legal information retrieval using topic clustering and neural networks. In Proc. 4th ICAIL Competition on Legal Information Extraction and Entailment (COLIEE), volume 47 of EPiC Series in Computing, pages 68–78.

[71] Newman, M. E. J. (2002). Assortative mixing in networks. Physical Review Letters, 89(20).

[72] Newman, M. E. J. (2003). Mixing patterns in networks. Physical Review E, 67(2).

[73] Nguyen, H., Nguyen, P. M., Vuong, T., Bui, Q. M., Nguyen, C. M., Dang, T. B., Tran, V., Nguyen, M. L., and Satoh, K. (2021). JNLP team: Deep learning approaches for legal processing tasks in COLIEE 2021. CoRR, abs/2106.13405.

[74] Nguyen, T., Nguyen, L., Tojo, S., Satoh, K., and Shimazu, A. (2017). Single and multiple layer BI-LSTM-CRF for recognizing requisite and effectuation parts in legal texts. In Proc. ICAL Workshops.

[75] Nguyen, T., Nguyen, L., Tojo, S., Satoh, K., and Shimazu, A. (2018). Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. Artif. Intell. Law, 26(2):169–199.

[76] O'Neill, J., Buitelaar, P., Robin, C., and O'Brien, L. (2017). Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), pages 159–168.

[77] OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2023). Gpt-4 technical report.

[78] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Proc. NAACL-HLT, pages 2227–2237.

[79] Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In Proc. 6th Italian Conf. on Computational Linguistics, volume 2481 of CEUR Workshop Proceedings. CEUR-WS.org.

[80] Puccinelli, D., Demartini, S., and D'Aoust, R. E. (2019). Fixing comma splices in italian with BERT. In Proc. 6th Italian Conf. on Computational Linguistics, volume 2481 of CEUR Workshop Proceedings. CEUR-WS.org.

[81] Rabelo, J., Kim, M., and Goebel, R. (2019). Combining similarity and transformer methods for case law entailment. In Proc. Int. Conf. on Artificial Intelligence and Law (ICAIL), pages 290–296.

[82] Rabelo, J., Kim, M., Goebel, R., Yoshioka, M., Kano, Y., and Satoh, K. (2020). COLIEE 2020: Methods for legal document retrieval and entailment. In Proc. of the JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, volume 12758 of New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, pages 196–210. Springer.

[83] Radford, A. and Sutskever, I. (2018). Improving language understanding by generative pre-training. In arxiv.

[84] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

[85] Rahat, T. A., Long, M., and Tian, Y. (2022). Is your policy compliant?: A deep learning-based empirical study of privacy policies' compliance with GDPR. In Proc. of the 21st Workshop on Privacy in the Electronic Society (WPES 2022), pages 89–102. ACM.

[86] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pages 1135–1144. ACM.

[87] Sakhaee, N. and Wilson, M. C. (2021). Information extraction framework to build legislation network. Artificial Intelligence and Law, 29(1):35–58.

[88] Sanchez, L., He, J., Manotumruksa, J., Albakour, D., Martinez, M., and Lipani, A. (2020). Easing legal news monitoring with learning to rank and BERT. In Proc. ECIR, volume 12036 of Lecture Notes in Computer Science, pages 336–343. Springer.

[89] Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., and Ma, S. (2020). BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In Proc. IJCAI, pages 3501–3507.

[90] Simeri, A. and Tagarelli, A. (2023a). Exploring domain and task adaptation of LamBERTa models for article retrieval on the Italian Civil Code. In Proc. of the 19th Conference on Information and Research science Connecting to Digital and Library science (IRCDL 2023), volume 3365 of CEUR Workshop Proceedings, pages 130–143.

[91] Simeri, A. and Tagarelli, A. (2023b). GDPR Article Retrieval based on Domain-adaptive and Task-adaptive Legal Pre-trained Language Models. In Proceedings of the 1st Legal Information Retrieval meets Artificial Intelligence Workshop LIRAI 2023 co-located with the 34th ACM Hypertext Conference HT 2023, volume 3594 of CEUR Workshop Proceedings, pages 63–76.

[92] Sovrano, F., Palmirani, M., and Vitali, F. (2020). Legal knowledge extraction for knowledge graph based question-answering. In Legal Knowledge and Information Systems, pages 143–153. IOS Press.

[93] Sulea, O., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., and van Genabith, J. (2017). Exploring the use of text classification in the legal domain. In Proc. ICAIL Workshop on Automated Semantic Analysis of Information in Legal Texts.

[94] Sulis, E., Humphreys, L., Vernero, F., Amantea, I. A., Di Caro, L., Audrito, D., and Montaldo, S. (2020). Exploring network analysis in a corpus-based approach to legal texts: A case study. In Proc. of the CAiSE for Legal Documents (COUrT) Workshop co-located with the the 33rd International Conference on Advanced Information Systems, pages 27–38.

[95] Surden, H. (2019). Artificial intelligence and law: An overview. 35 GA. ST. U. L. REV., 1305.

[96] Tagarelli, A. and Simeri, A. (2022a). LamBERTa: Law Article Mining Based on Bert Architecture for the Italian Civil Code. In Proc. the 18th Italian Research Conference on Digital Libraries, volume 3160 of CEUR Workshop Proceedings. CEUR-WS.org.

[97] Tagarelli, A. and Simeri, A. (2022b). Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. Artif. Intell. Law, 30(3):417–473. Published: 15 September 2021.

[98] Torre, D., Abualhaija, S., Sabetzadeh, M., Briand, L. C., Baetens, K., Goes, P., and Forastier, S. (2020). An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR. In Proc. of the 28th IEEE International Requirements Engineering Conference (RE 2020), pages 136–146. IEEE.

[99] van der Maaten, L. and Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9:2579–2605.

[100] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Proc. NIPS, pages 5998–6008.

[101] Viola, L. (2017). Interpretazione della legge con modelli matematici, volume 26. Diritto Avanzato.

[102] Wang, P., Yang, Z., Niu, S., Zhang, Y., Zhang, L., and Niu, S. (2018). Modeling dynamic pairwise attention for crime classification over legal articles. In Proc. ACM SIGIR, pages 485–494.

[103] Wang, S., Khabsa, M., and Ma, H. (2020). To pretrain or not to pretrain: Examining the benefits of pretraining on resource rich tasks. In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 2209–2213. ACL.

[104] Whalen, R. (2016). Legal networks: The promises and challenges of legal network analysis. Michigan State Law Review, page 539.

[105] Wiggers, G. and Verberne, S. (2019). Citation metrics for legal information retrieval systems. In Proc. of the 8th International Workshop on Bibliometric-enhanced Information Retrieval co-located with the 41st European Conference on Information Retrieval, volume 2345 of CEUR Workshop Proceedings, pages 39–50. CEUR-WS.org.

[106] Yamakoshi, T., Komamizu, T., Ogawa, Y., and Toyama, K. (2019). Japanese mistakable legal term correction using infrequency-aware BERT classifier. In Proc. IEEE Int. Conf. on Big Data, pages 4342–4351.

[107] Yang, W., Jia, W., Zhou, X., and Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. In Proc. IJCAI, pages 4085–4091.

[108] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., and Hovy, E. H. (2016). Hierarchical attention networks for document classification. In Proc NAACL-HLT, pages 1480–1489. The Association for Computational Linguistics.

[109] Ye, H., Jiang, X., Luo, Z., and Chao, W. (2018). Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In Proc. NAACL-HLT, pages 1854—1864.

[110] Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. CoRR, abs/1702.01923.

[111] Yoshioka, M., Aoki, Y., and Suzuki, Y. (2021). BERT-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task. In Proc. of the 18th International Conference for Artificial Intelligence and Law (ICAIL), pages 278–284. ACM.

[112] Zhang, B., Ding, D., and Jing, L. (2023). How would stance detection techniques evolve after the launch of chatgpt?

[113] Zhang, P. and Koppaka, L. (2007). Semantics-based legal citation network. In Proc. of the 11th International Conference on Artificial Intelligence and Law, ICAIL '07, pages 123–130.

[114] Zhao, Y. and Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. Mach. Learn., 55(3):311–331.

[115] Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. (2021). When does pretraining help? assessing self-supervised learning for law and the casehold dataset.

[116] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In Proc. ACL.

[117] Zhou, X., Zhang, Y., Liu, X., Sun, C., and Si, L. (2019). Legal intelligence for e-commerce: Multi-task learning by leveraging multiview dispute representation. In Proc. ACM SIGIR, pages 315–324.

[118] Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. pages 45–50.